4

RADC-TR-88-11, Vol VII (of eight)
Interim Technical Report
June 1988

NORTHEAST ARTIFICIAL INTELLIGENCE CONSORTIUM ANNUAL REPORT 1986 Artificial Intelligence Applications to Speech Recognition

Syracuse University

H. E. Rhody, J. Hillenbrand and J. A. Biles

This effort was funded partially by the Laboratory Director's fund.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss AFB, NY 13441-5700

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-88-11, Vol VII (of eight) has been reviewed and is approved for publication.

APPROVED: John G. Parker

JOHN G. PARKER
Project Engineer

APPROVED: Garry W. Barringer

GARRY W. BARRINGER
Technical Director
Directorate of Intelligence & Reconnaissance

FOR THE COMMANDER: James W. Hyde III

JAMES W. HYDE III
Directorate of Plans & Programs

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441-5700. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS
N/A

2a. SECURITY CLASSIFICATION AUTHORITY
N/A

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE
N/A

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
N/A

5. MONITORING ORGANIZATION REPORT NUMBER(S)
RADC-TR-88-11, Vol VII (of eight)

6a. NAME OF PERFORMING ORGANIZATION
Northeast Artificial Intelligence Consortium (NAIC)

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Rome Air Development Center (COES)

6c. ADDRESS *(City, State, and ZIP Code)*
409 Link Hall
Syracuse University
Syracuse NY 13244-1240

7b. ADDRESS *(City, State, and ZIP Code)*
Griffiss AFB NY 13441-5700

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Rome Air Development Center

8b. OFFICE SYMBOL *(If applicable)*
COES

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F30602-85-C-0008

8c. ADDRESS *(City, State, and ZIP Code)*
Griffiss AFB NY 13441-5700

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.
62702F	5581	27	13

11. TITLE *(Include Security Classification)*
NORTHEAST ARTIFICIAL INTELLIGENCE CONSORTIUM ANNUAL REPORT 1986 Artificial Intelligence Applications to Speech Recognition

12. PERSONAL AUTHOR(S)
H. E. Rhody, J. Hillenbrand, J. A. Biles

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM Jan 86 TO Dec 86

14. DATE OF REPORT *(Year, Month, Day)*
June 1988

15. PAGE COUNT
72

16. SUPPLEMENTARY NOTATION
This effort was performed as a subcontract by RIT Research Corporation to Syracuse University, Office of Sponsored Programs. (See reverse)

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
23	02	
05	07	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Artificial Intelligence, Speech Recognition, Expert Systems, Signal Processing, Phoneme Classification, Knowledge-based Systems.

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The Northeast Artificial Intelligence Consortium (NAIC) was created by the Air Force Systems Command, Rome Air Development Center, and the Office of Scientific Research. Its purpose is to conduct pertinent research in artificial intelligence and to perform activities ancillary to this research. This report describes progress that has been made in the second year of the existence of the NAIC on the technical research tasks undertaken at the member universities. The topics covered in general are: versatile expert system for equipment maintenance, distributed AI for communications system control, automatic photo interpretation, time-oriented problem solving, speech understanding systems, knowledge base maintenance, hardware architectures for very large systems, knowledge-based reasoning and planning, and a knowledge acquisition, assistance, and explanation system. The specific topic for this volume is the design and implementation of a knowledge-based system to read speech spectrograms.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
John G. Parker

22b. TELEPHONE *(Include Area Code)*
(315) 330-4024

22c. OFFICE SYMBOL
RADC (IRAA)

Previous editions are obsolete.

Block 10 (Cont'd)

Program Element Number	Project Number	Task Number Number	Work Unit Number
62702F	4594	18	E2
61101F	LDFP	15	C4
61102F	2304	J5	01
33126F	2155	02	10

Block 16 (Cont'd)

This effort was funded partially by the Laboratory Directors' Fund.

7 ARTIFICIAL INTELLIGENCE APPLICATIONS TO SPEECH RECOGNITION

Report submitted by:

Harvey E. Rhody
James Hillenbrand
John A. Biles

RIT Research Corporation
75 Highpower Road
Rochester, New York 14623

TABLE OF CONTENTS

7.1 System Architecture

Speech understanding can be loosely defined as natural language understanding with noisy or uncertain input. In traditional speech recognition the goal is to derive words from speech signals, while the goal of a speech understanding system is to understand the intent or meaning of a spoken utterance. Clearly speech recognition is a component of speech understanding, but recognition is only the "front end" of the system; natural language understanding also must be done on the recognized words in order to represent the intent of the utterance.

The guiding philosophy of RIT's Speech Understanding project centers on the contention that it is possible for humans to reliably "read" speech spectrograms. Since this is a cognitive process and can be "explained" by the person doing it, we believe it possible to build a knowledge based system containing that expertise. The overall goal, then is to design and implement a knowledge based system that reads speech spectrograms.

We view the architecture of this system on two levels, a "virtual" architecture at the software level and a "physical" architecture at the hardware level. Most of the work to date has concentrated on the software architecture, but the acquisition of two TI Explorers with additional signal processing hardware has provided us with a good hardware configuration onto which to map our software architecture.

7.1.1 Software Architecture

The software architecture of the system is highly modular and largely data-driven. Besides the usual software design and management concerns, a high degree of modularity was desirable for other reasons as well. One very pragmatic reason is the fact that there is a high degree of turnover among the graduate students working on the project. Graduate students at RIT are required to complete a Masters thesis, the scope of which is modest compared to a typical dissertation. Since a typical student will spend at most a year involved in the project, a highly modular system architecture allows us to break off relatively small, isolated chunks of the system for Masters theses.

Another reason for emphasizing small, relatively independent modules is the current state of flux that characterizes our available hardware. We are in the process of converting to the TI Explorer Lisp environment from the Sun UNIX environment. While this conversion is taking place, we still have thesis work in progress that will have to be converted at a later time. Keeping the individual projects small will ease their eventual conversion.

The decision to use a data-driven control strategy was made under the assumption that reliable feature extraction could be done in parallel at a low level, eliminating the need for more sophisticated (e.g. blackboard) control strategies. We feel that the goal of real time performance precludes the use of exotic control strategies and that a largely data-driven approach can allow enough flexibility, provided that reliable feature extraction can be done.

The software architecture is summarized by Figure 7-1. A spoken utterance is digitized by an analog-digital converter to produce a digitized "oscillogram" of the raw speech sample. The digitized raw speech is then passed through a fast Fourier transform to produce a speech spectrogram. This represents the typical "input" to a human expert, who can "read" a speech spectrogram, and it can be thought of as the input to our knowledge based system.

The spectrogram is then given to a spectral segmentation module whose job it is to partition the speech spectrogram into relatively homogeneous segments. This is the first step taken by every human expert we have observed reading a spectrogram, and it basically amounts to drawing vertical lines on the spectrogram where changes occur. Spectral segmentation is discussed in more detail in Section 7-2.

The segmented spectrogram is then fed in parallel to a collection of low level, algorithmic feature extractors. Each of these modules computes a value or set of values for a specific feature over the span of the given spectral segment. These modules are capable of computing their values independently of one another, which means that a high degree of fine grain parallelism is possible. This inherent parallelism will be exploited when the software architecture is mapped onto a suitable hardware architecture that provides parallel processors. The work currently being done on these low level feature extractors is covered in Section 7-2.

After the feature extractors have computed their values, the continuous spectrogram has essentially been replaced by a discrete sequence of "feature vectors," one vector for each spectral segment. These vectors can be thought of as frames in an AI sense, with slots being features. This representation is more compressed than the original spectrogram by orders of magnitude, which means that the enormous amount of data in the original speech signal has been reduced to a manageable level, without sacrificing useful information. It is also important to note that after the feature vector has been generated for a segment, the FFT spectrogram and the original raw speech are no longer needed.

The spectral segment vectors are then given to a knowledge based "phoneme" builder, which joins segments to build intermediate level speech structures. This module can be thought of as doing knowledge based syntactic pattern recognition. The problem is syntactic or structural pattern recognition in that small segments of speech are being "parsed" to generate "phonemes." The module is

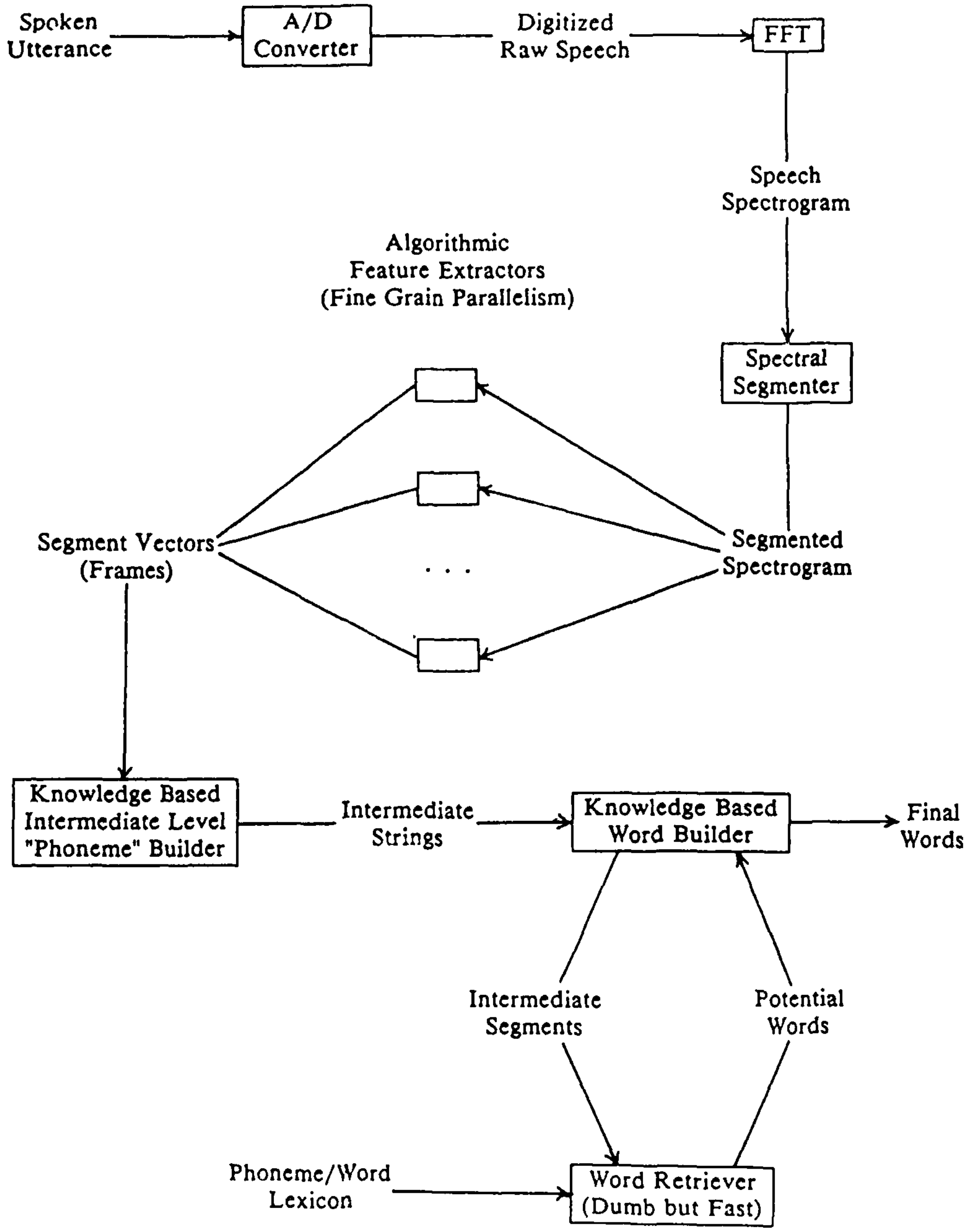


Figure 7-1. Software Architecture

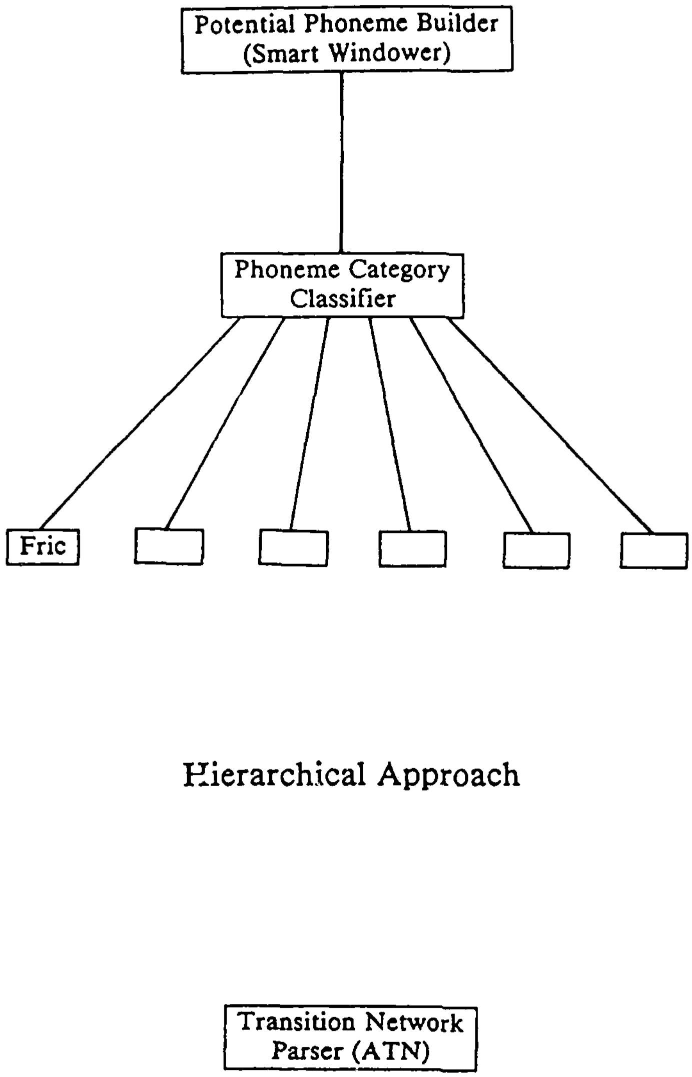
knowledge based in that the approach is to "clone" experts who can successfully perform this task.

We enclose "phoneme" in quotes because the intermediate level structures may or may not be actual phonemes in the linguistic sense. Some traditional phonemes (e.g. fricatives, stops) seem to be good candidates for intermediate level structures because they are readily recognizable as classes and, therefore, provide reliable categories for classification. Other traditional phoneme classes (e.g. vowels, glides) are harder to pin down and lead to less reliable classification. Whether we end up with phonemes, diphones, other units, or a combination of all of these will be determined by what units allow us to make the most reliable identification.

We are currently exploring two approaches to the design of the knowledge based "phoneme" recognizer. The first approach is hierarchical in nature and presupposes that we can identify the category or class of a given spectral segment with a high degree of reliability. Provided this can be done, a tree-like structure can be built with the leaf nodes representing specific "phonemes" and internal nodes representing categories of phonemes that are less specific the closer they are to the root of the tree (see Figure 7-2).

The second approach to the "phoneme" recognizer is to build an augmented transition network (ATN) that uses spectral segments as inputs to fire transitions among states. This approach is appealing in that it bundles together the problem of deciding the boundaries between "phonemes" with the problem of identifying the actual "phonemes." The hierarchical approach assumes that setting inter-phoneme boundaries can be done reliably without knowing yet the identities of the phonemes. While human experts often seem to be able to do this, it seems more likely that drawing boundaries and identifying "phonemes" are inextricably linked. The work already done on a hierarchically based fricative identifier (covered in Section 7.4) can be incorporated easily into an ATN based parser since it provides knowledge on how to construct the states in the ATN that pertain to parsing fricatives.

The question of what goes in the algorithmic feature extractors and what goes in the knowledge based recognizers boils down to whether a given task can be done well with a "standard" algorithm or whether it requires AI techniques. Our philosophy is that when a knowledge based recognizer needs to know something about the spectrogram or the raw speech, this interface to the spectrogram will be done in one of two ways. If the question about the spectrogram can be answered algorithmically, a new feature extraction module is created and added to the set of parallel feature extractors, and a new slot is created for the feature in the segment vectors. From a performance point of view, the calculation of this new feature is "free" in that it is computed in parallel with all the other algorithmically computed features.



Hierarchical Approach

Syntactic Approach

Figure 7-2. Alternative Approaches to Intermediate Level "Phoneme" Builder

If the question about the spectrogram requires a knowledge based approach, an extension is made to the existing knowledge based recognizer to ask this lower level question. Eventually, this extension will lead to reasonable "algorithmic" extractors that will deal directly with the spectrogram. This approach blurs the typical distinction between feature extraction, which is traditionally algorithmic, and recognition, which is usually AI based. We may have some "smart" feature extractors and some "dumb" recognizers, depending on whether a good algorithm can be developed to perform the task.

This data-driven organization means that higher level processes do not demand data from lower level processes. Instead, the lower level processes always comnute everything they know how to compute for every speech sample. This obviously requires a great deal of low level computation, but since that will be done in parallel, it costs processors, not time. The upshot is that high level processes always have all the data they need to make decisions, which obviates the need for a direct interface with the low level feature extractors.

The output of the knowledge based "phoneme" builder is a string of intermediate level structures -- some combination of phonemes, diphones, syllables, or other units. These intermediate strings are given to a "word builder," which tries to partition the intermediate string into segments that might be words, and then looks up the potential words in a lexicon of word/intermediate-string pairs. This module is largely knowledge based, and its organization has not yet been fixed. A simple generate-and-test strategy will be tried initially, but we anticipate that this approach will not provide anything approaching real-time performance.

A severe problem for the word builder is the uncertainty inherent in the intermediate strings it gets from the phoneme builder. Some "phonemes" in a typical intermediate string will be less certain than others. For example, a typical string might contain phonemes like "f sound" followed by "a glide" followed by "something vowel-like" followed by "not a stop." A pure generate-and-test strategy would have to try each of the words or word segments that fits this very rough outline and eliminate those words that do not fit with other words in the utterance that have already been recognized. Moreover, the problem of segmenting the intermediate string is probably as difficult, albeit on a different level, as the spectral-segmentation problem.

An obvious component of the system is a lexicon that maps intermediate string segments onto words. At the core of this lexicon is a fast retrieval module that can look up a given intermediate string and return any word(s) that match it. This retrieval module will be discussed in Section 7.5.

Finally at the highest level of the system, a knowledge based component that does traditional natural language understanding will derive the intent of the utterance. This clearly will have to be a domain-specific system that has sufficient knowledge of the domain of discourse to be able to make sense of an utterance. It's job will be made more difficult than that of traditional natural language understanding systems because the words being given to it may not be correct. Understanding natrual language with uncertain words is a largely untapped area in natuaral language research.

7.1.2 Hardware Architecture

The NAIC was instrumental in helping RIT obtain two TI Explorer Lisp machines, one of which is equipped with TI's Odyssey signal processing board. Briefly, the Odyssey board consists of four TMS 32020 processors and on-board program and data memory connected with an on-board bus. The board talks to the Explorer Lisp environment at a hardware level via the Explorer's Nu Bus. At a software level, access is done via a memory mappea protocol, with TI-supplied software to facilitate up-loading, down-loading and processor control. A D-A/A-D device is being designed by TI to sit on the Odyssey board and handle speech input and digitization.

With this hardware configuration, it is possible to map the software architecture onto a hardware architecture that allows the exploitation of the parallelism inherent in the four TMS 32020 processors. Since these processors were designed especially for doing low level number crunching in signal processing environments, they are ideal for serving as the hardware for computing FFTs and doing algorithmic based feature extraction. The Explorer/Odyssey, then, provides an ideal hardware configuration onto which to map the software architecture because it supplies state-of-the-art processing for both low level signal processing and high level AI programs. During the coming year we will be moving the project from the Sun system to the Explorer/Odyssey machine.

7.1.3 Future Directions

When the project has been moved to the TI Explorers, we will finally have a suitable environment for doing real knowledge based system development. Up until now our AI tools have been limited to RuleMaster, an induction based expert system building tool, and while this is a good tool for a UNIX environment and provides an excellent testbed for developing phoneme recognizers, it is rather impoverished when compared to full-blown tools that exist for the Explorer environment. We plan to move more heavily into the higher level recognition modules as better tools become available to us.

We also plan to begin exploiting the parallelism provided by the Odyssey board on the Explorer to map our parallel software architecture onto the Odyssey's TMS 32020 processors. This hinges in part on the development of suitable tools for processing speech signals.

7.2 Analysis/Display Tools and Signal Processing Software

7.2.1 SpeechTool

A speech analysis software package called SpeechTool has been developed for our Sun 2/130 workstation. The package performs a variety of different types of spectral analysis and has the capability to graphically display the results of all of the commonly used speech-analysis techniques. Graphic capabilities include the following displays, shown in Figures 7-3 to 7-13:

1. FFT spectrograms (Fig. 7-3b, 7-4a, 7-5e, 7-6e, 7-7c, etc.)
2. LPC spectrograms (Fig. 7-3a)
3. Oscillograms (Fig. 7-5d)
4. User-defined measurement functions: For lack of a better term, a mixed category of derived waveforms that constitute one measurement per millisecond. Examples include rate of spectral change (Fig. 7-5a), total energy per 10-msec speech segment (Fig. 7-5b, 7-5c) and zero-crossing rate (Fig. 7-8b).
5. Individual FFT spectra for specified frames (Fig. 7-11).
6. 'Waterfall' spectral displays: These displays show how the LPC or FFT spectrum evolves over time (Fig. 7-12, 7-13).
7. Time (Fig. 7-6d) and frequency markers (Fig. 7-13).

7.2.2 Calculating an Auditory Spectrum

As can be seen from the displays in Figures 7-3 to 7-13, speech analysis consists primarily of determining how the speech spectrum changes as a function of time. The problem of automatic phonetic recognition is largely one of relating those patterns of spectral change to specific phonemes, diphones, syllables, etc. Because spectral analysis plays such a central role in speech recognition research, it is very important that methods be found to represent the spectrum in a way that will optimize the ability to accurately categorize the patterns of spectral change.

Most speech recognition systems use linear Fourier spectra or Linear Predictive Coding spectra. Although these kinds of analyses correspond roughly to the kinds of signal analyses performed by the human auditory system, the correspondence is only approximate. The goal of our spectral analysis work was to develop and test methods for producing spectral representations that corresponded more closely to the nonlinear spectral representations that are provided by the human auditory system.

Spectral representations of speech signals are chosen over time-domain representations because this type of analysis approximates the kind of signal processing performed by the peripheral auditory system of humans. The mechanical and neural action of the basilar membrane and other cochlear structures function to direct high-frequency signal components to the basal end of the cochlea and low-frequency signal components to the

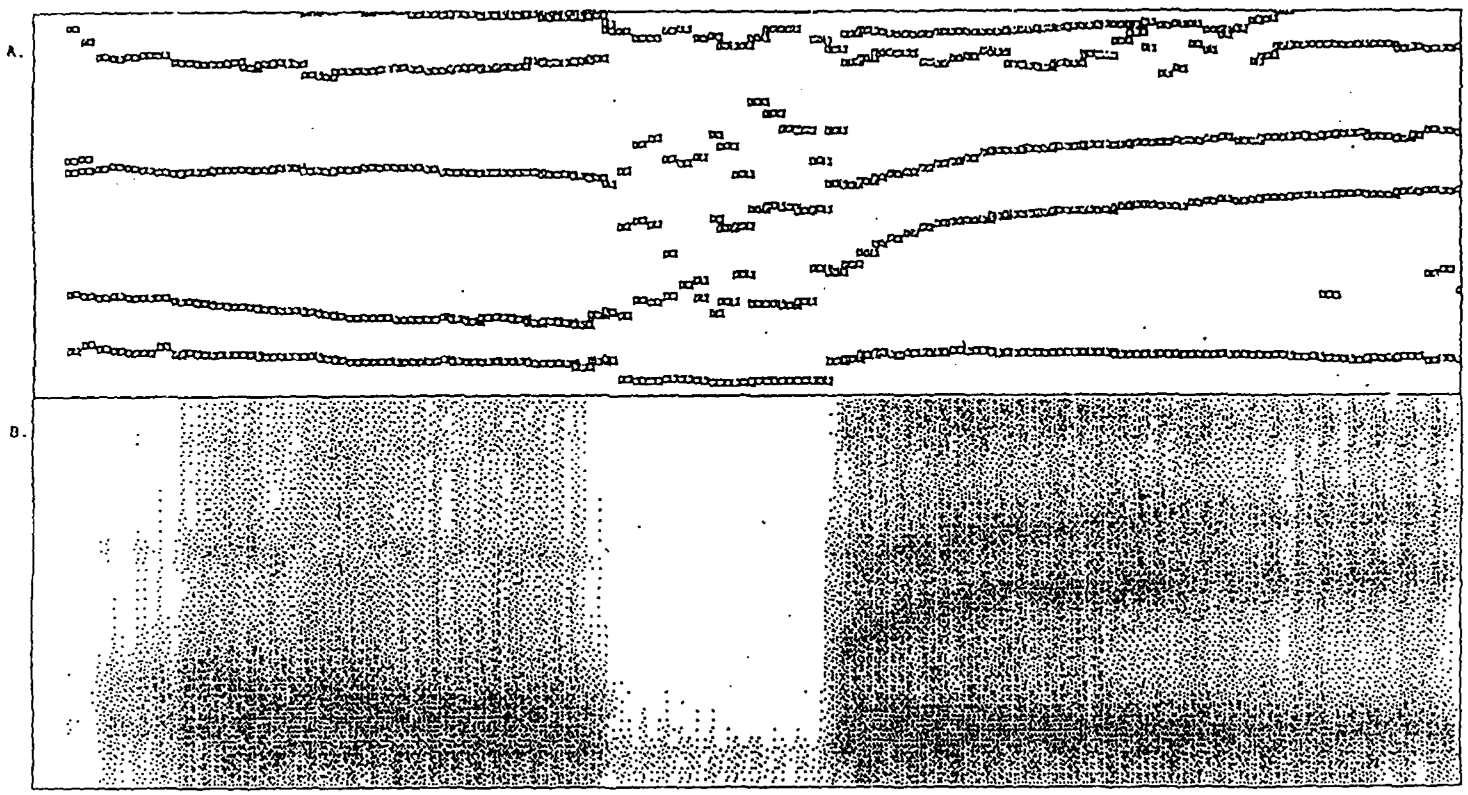


Figure 7-3 Output from a standard LPC formant tracker (a) and FFT spectrogram (b) for the word "obey"

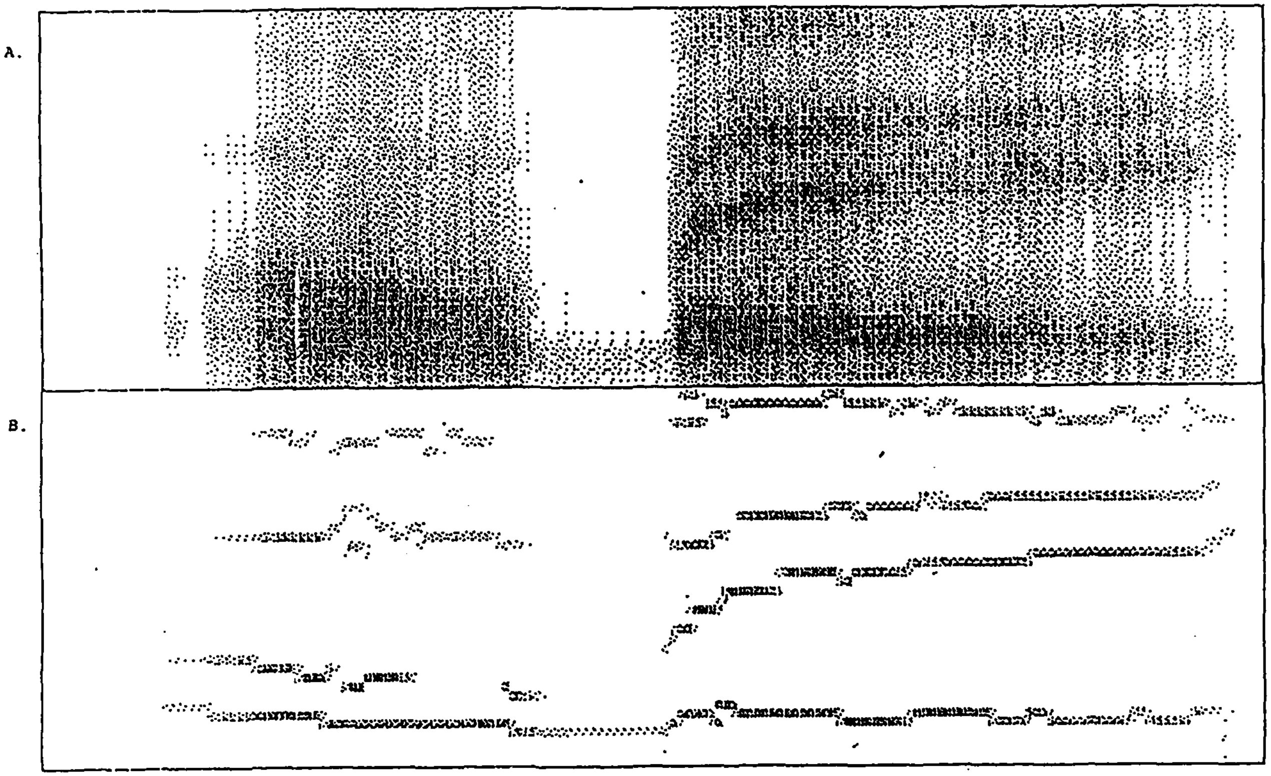


Figure 7-4. Spectrogram of the word "obey" (a) and output of an early version of the formant tracker described in Section 7.2.2.

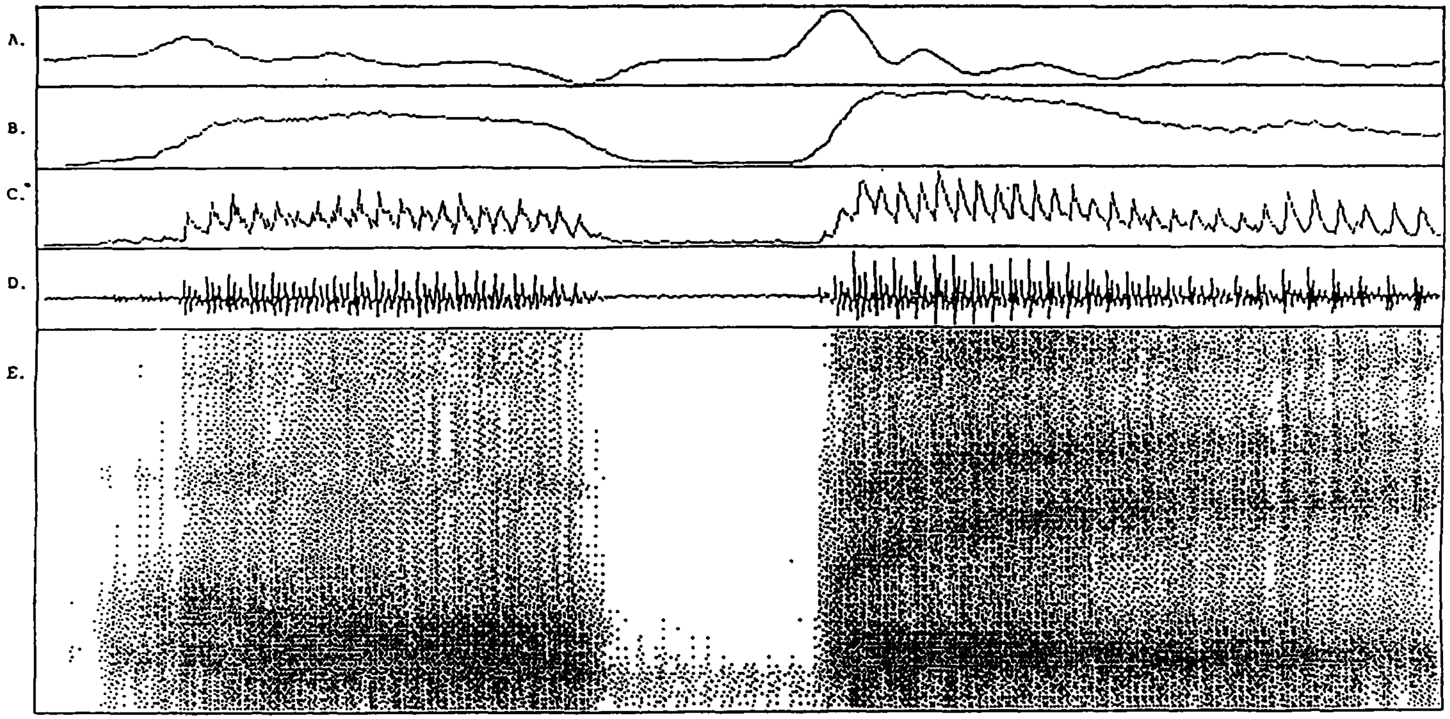


Figure 7-5. Spectrogram with several time-aligned measurement functions: (a) rate of spectral change ("spectral derivative"), (b) smoothed energy function, (c) unsmoothed energy function, (d) original digital waveform, and (e) spectrogram.

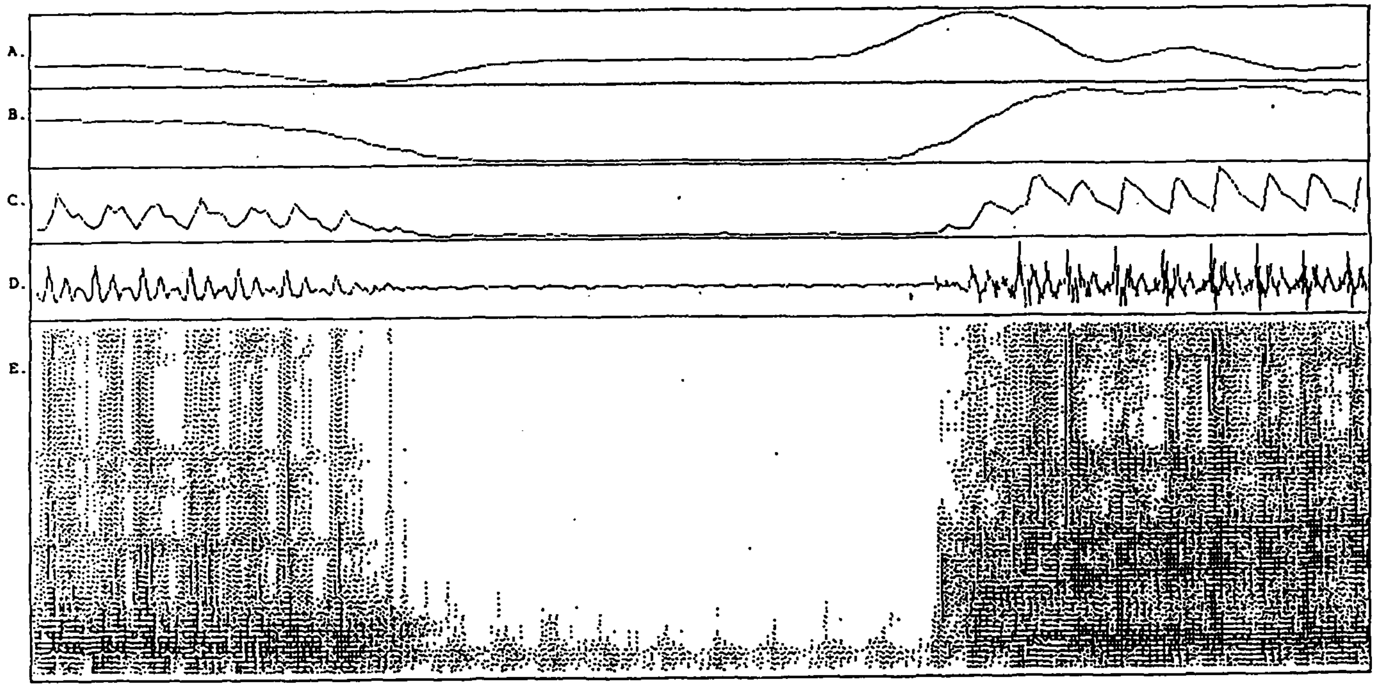


Figure 7-6. Expanded view of the functions depicted in Figure 7-5.

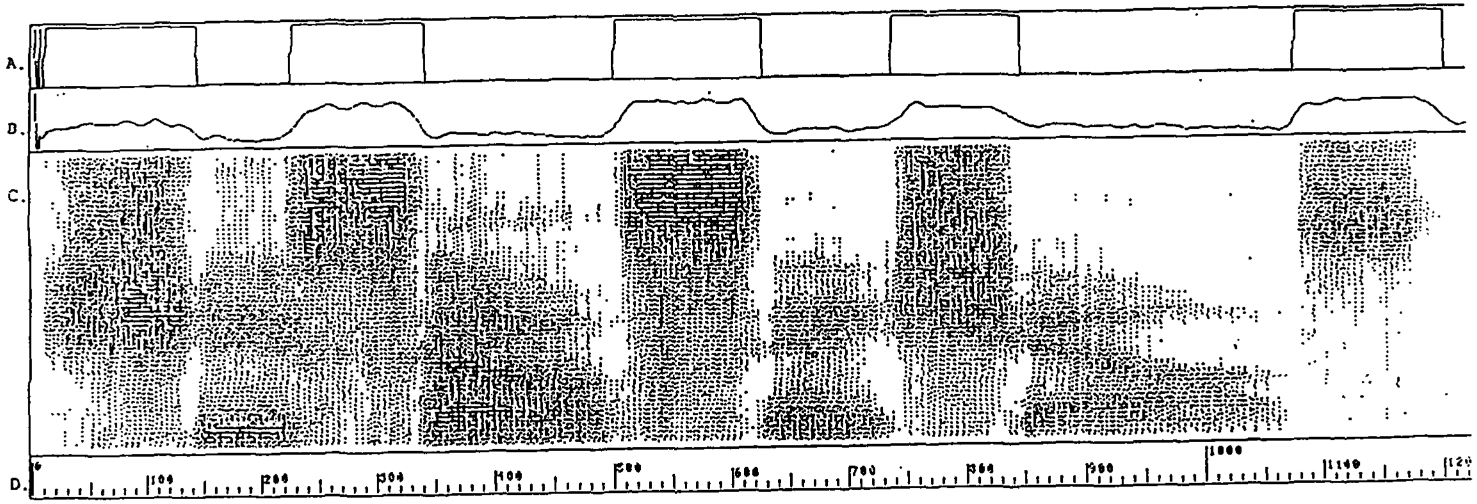


Figure 7-7. Spectrogram and derived measurements of the utterance "she sells sea shells". (a) fricative detector based on (b) zero-crossing count, FFT spectrogram, (d) time scale.

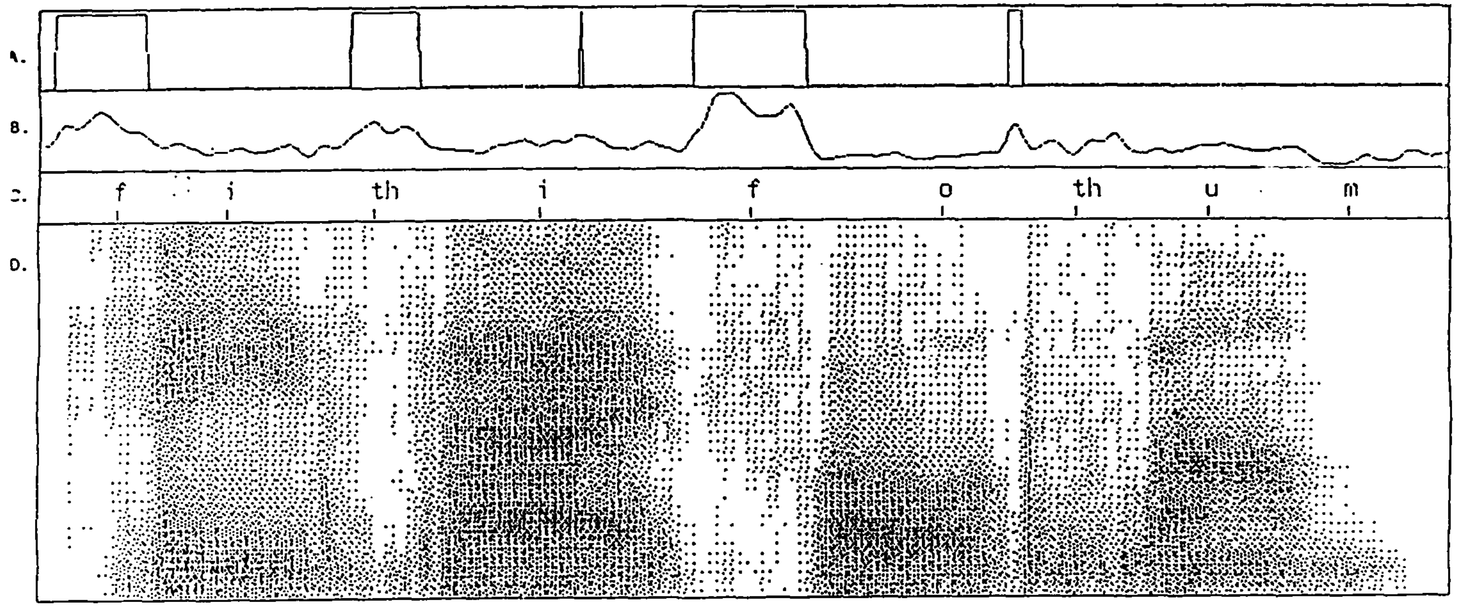


Figure 7-8. Spectrogram and derived measurements of the utterance "fee thigh fo thumb". (a) fricative detector based on (b) zero-crossing count, (c) phonetic segment labels, (d) FFT spectrogram.

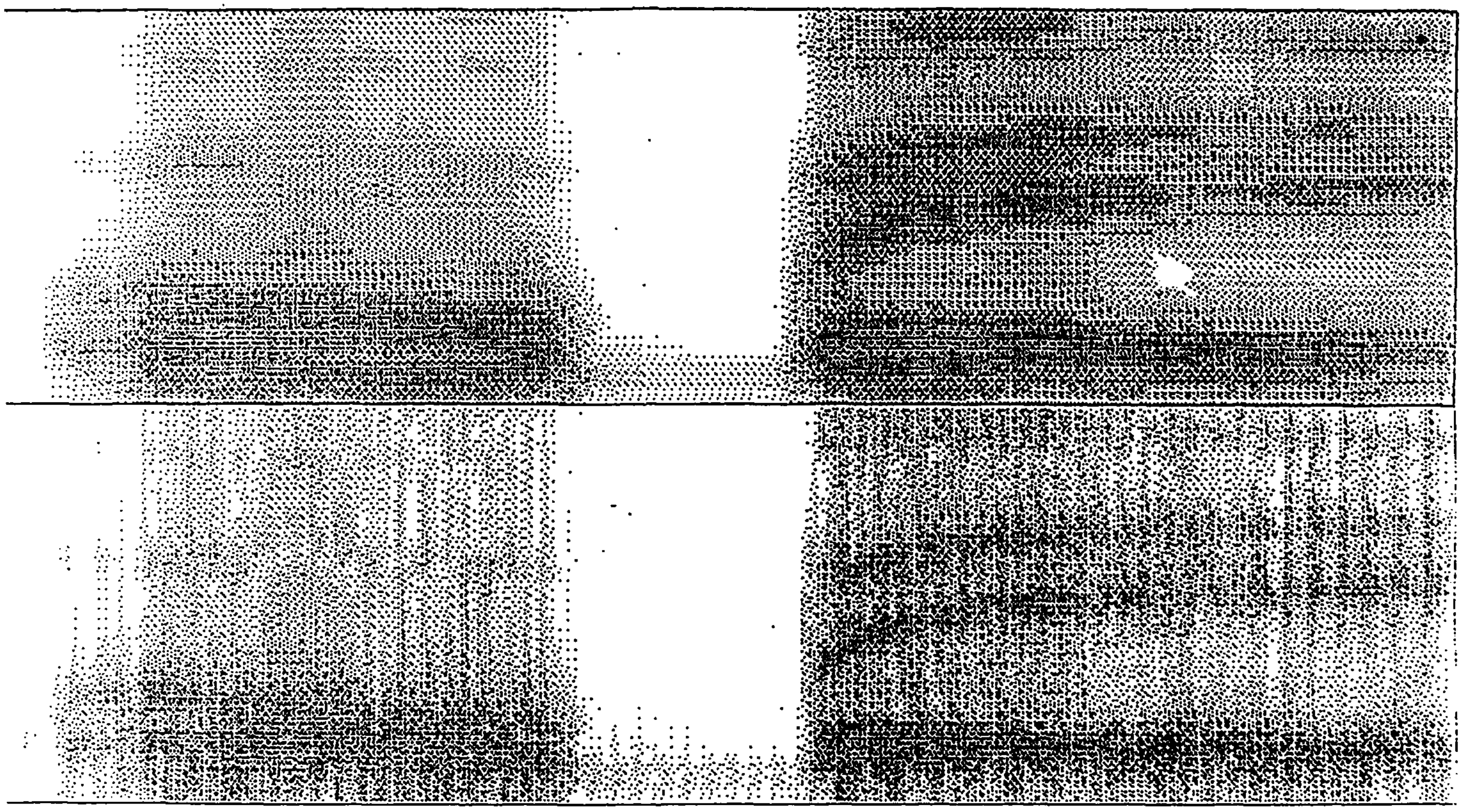


Figure 7-9. Smoothed (top) and unsmoothed (bottom) FFT spectrograms.

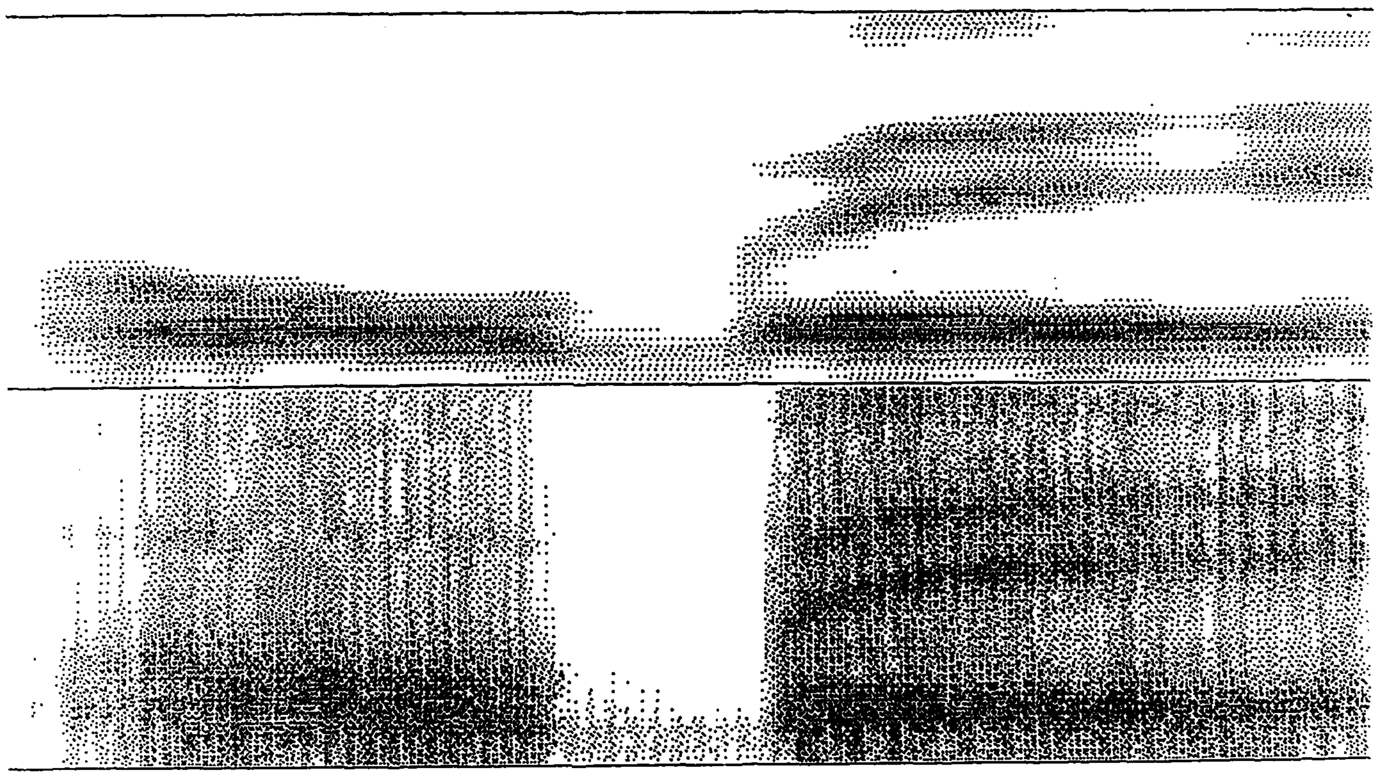


Figure 7-10. Comparison of a standard FFT spectrogram (bottom) with an FFT spectrogram designed to emphasize the formant regions (top).

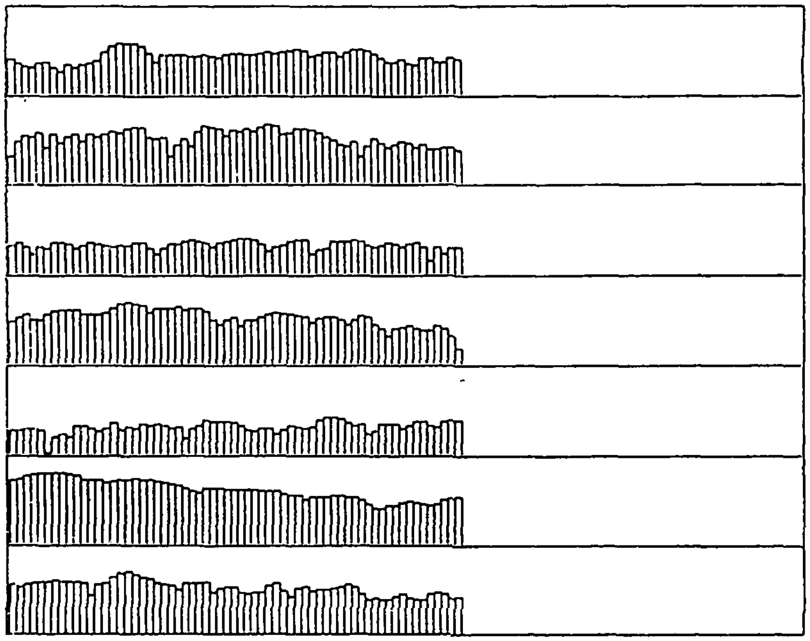


Figure 7-11. Individual FFT spectra from specified time slices of the signal.

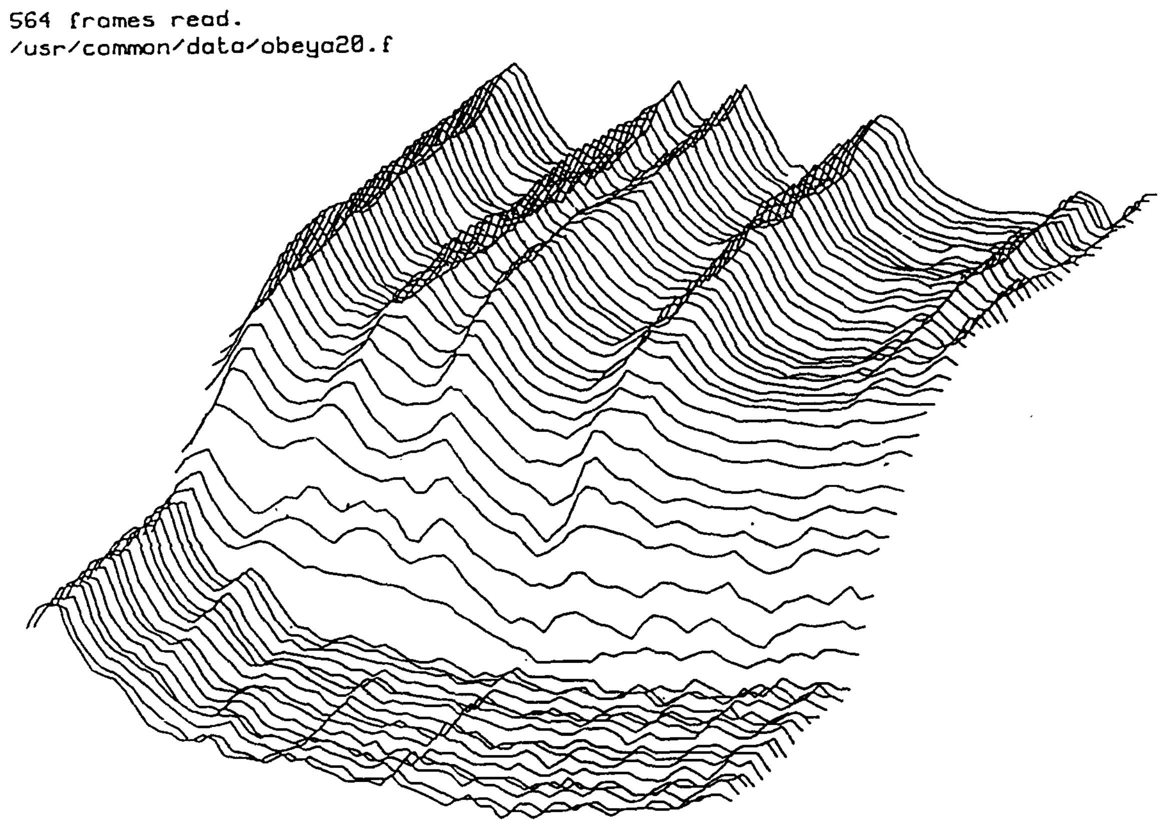


Figure 7-12. Waterfall display showing how the smoothed FFT spectrum evolves over time.

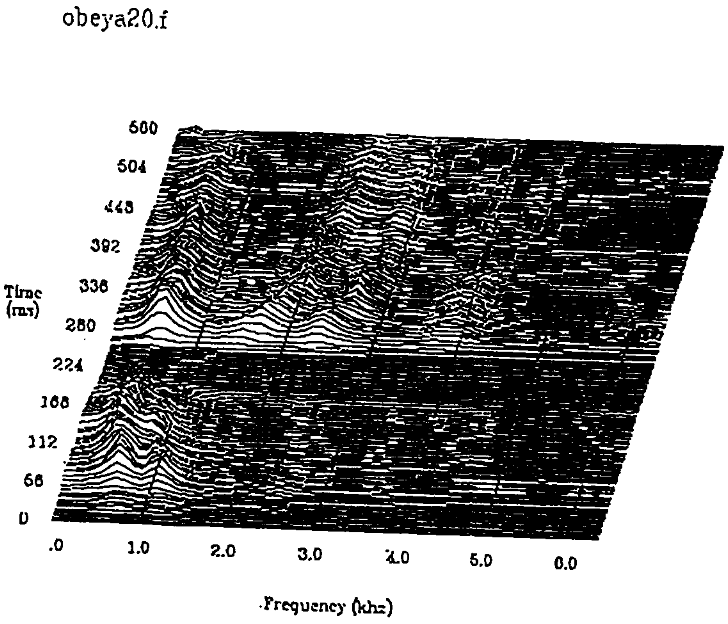


Figure 7-13. Waterfall display of a larger segment of the utterance shown in Figure 7-12.

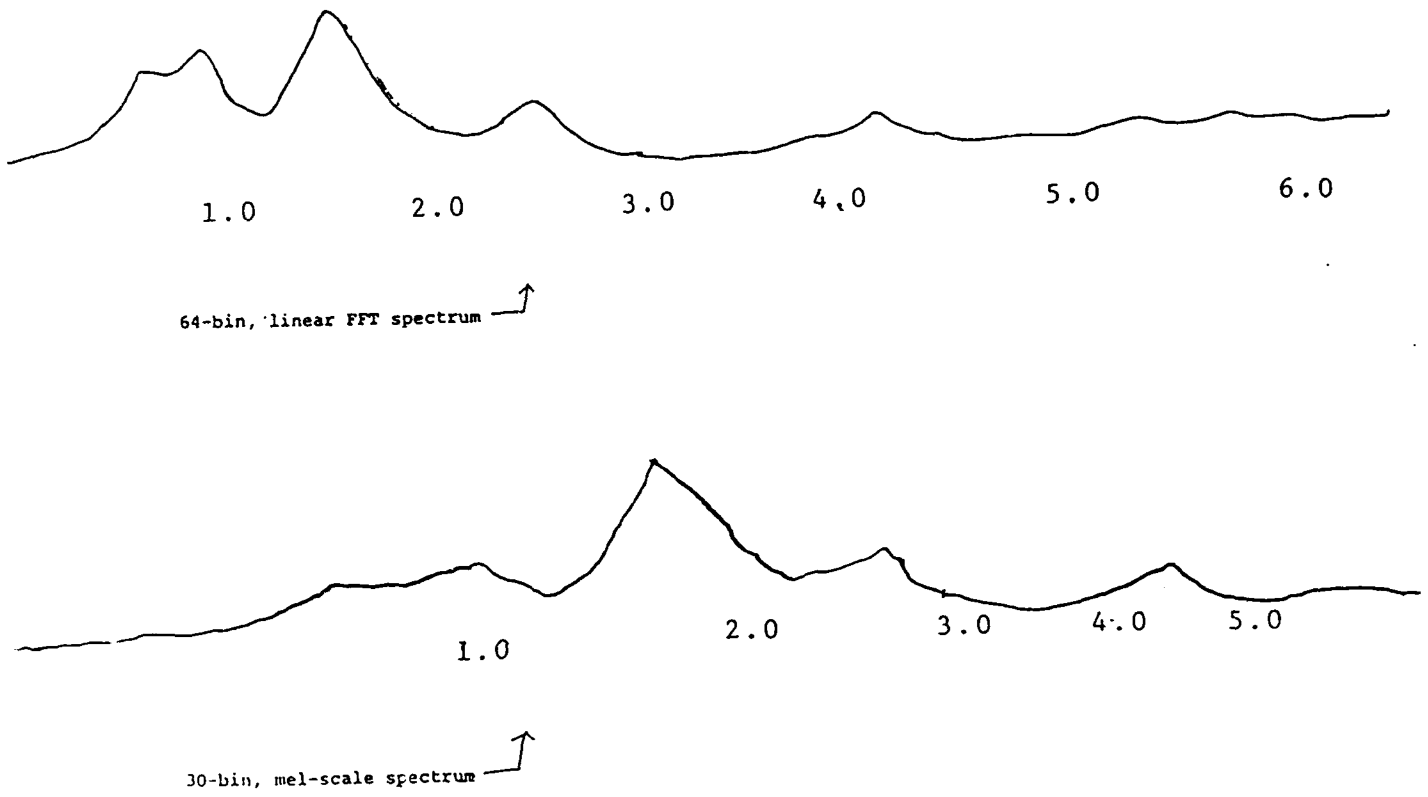


Figure 7-14. Comparison of a linear FFT with a nonlinear, auditory spectrum based on the mel scale.

apical end of the cochlea. Tonotopically arranged fibers of the auditory nerve then carry electrical signals that code the amount of basilar membrane motion at particular locations along the cochlea.

Power spectra calculated by FFT or LPC algorithms provide a rough approximation to this kind of frequency coding by producing an array of numbers whose magnitudes are proportional to signal intensities in particular frequency regions. However, the correspondence between FFT and LPC spectra and the kind of "neural spectrogram" provided by the peripheral auditory system is known to be only approximate. In a linear spectrum, frequency resolution is constant across the spectrum. For example, a 128-point FFT returns 64 magnitude points and 64 phase points. Assuming a sample frequency of 12.8 kHz (and therefore a signal bandwidth of 6.4 kHz), each of the 64 bins corresponds to 100 Hz. Despite the popularity of linear spectra such as these, it is well known that auditory spectra are non-linear; i.e., bandwidths are not constant across the spectrum. Specifically, it is well known that frequency resolution is better at low frequencies than at high frequencies. As a consequence, spectral distance measures that use linear spectra would be expected to underestimate the importance of differences in the low frequencies and, conversely, overestimate the importance of differences in the high frequencies.

Nonlinear spectra can be derived from linear spectra simply by the appropriate summing of adjacent frequency bins in the linear spectrum. To approximate the kind of frequency resolution in the auditory system one would sum a small number of bins in the low frequencies and a large number of bins in the high frequencies. The main problem is to determine exactly what bins should be summed to approximate an auditory spectrum. After experimenting with several different schemes, we have settled on a system based on the mel scale. This scale was empirically derived from psychoacoustic experiments involving pitch matching. A given distance along the mel scale corresponds to specific change in perceived pitch rather than a constant change in signal frequency. Figure 7-13 shows a mel-scale waterfall-type display of the word "obey" using a mel-scale transform of a linear FFT spectrogram. Figure 7-14 compares a 64-bin linear FFT of the midpoint of a vowel with a 30-bin mel-scale equivalent. Notice that frequency resolution is relatively good in the low frequencies and relatively gross in the high frequencies.

7.2.3 Evaluation

Diphone and word-recognition tests were conducted using a template-matching system based on spectral differences. The recognition tests compared the 64-bin linear FFTs with the 30-bin mel-scale equivalent spectra. The results showed consistently better performance for the mel-scale spectra. In addition, due to the data reduction involved in reducing the 64-bin spectra to 30 bins, calculations with the mel-scale spectra are considerably faster than the 64-bin linear spectra.

7.2.4 Formant Tracking

It has been known for many years that formant frequency patterns are one of the most important sources of phonetic information in the speech signal. For this reason, a major goal of our signal processing research is to develop algorithms to track formant frequencies. Formants appear on spectrograms as broad bands of energy. Although it is relatively easy for a trained phonetician to pick out formants visually, it has proven to be a difficult task to perform automatically. The errors that are commonly encountered include failing to detect the presence of a formant, failing to detect when two formants merge, and spurious detection of a formant where none exists. For example, if F1 and F2 merge together the correct F2 may be missed, with F3 being labeled as F2. Since recognition based on formants is very sensitive to this labeling, such errors can lead to serious mistakes. There have been many attempts at formant tracking, using a variety of techniques. We have investigated several of these in trying to determine the best approach for our purposes.

Most formant tracking methods in recent years have been based on picking peaks from linear prediction spectra. Linear prediction is a simple and powerful technique based on an all-pole model of the vocal tract. New coefficients for the model are computed approximately every 10-20 msec and used to generate a series of spectra. The peaks of these spectra are the raw data for formant tracking. The peaks are assigned to formants based on a set of rules designed to make the formant trajectories fairly continuous. A simple slot filling scheme due to Markel (1975) which we have implemented appears to do a reasonable job. It is claimed that this method is about 85% accurate, but there is no data available to support this. A more sophisticated algorithm due to McCandless (1974) uses 'anchor points' and more detailed rules, and claims to achieve accuracies as high as 90%, but once again without any documentation. This algorithm also uses spectral enhancement techniques for resolving the problem of formant mergers. These peak-picking methods are notable for the ease with which they achieve fairly good results, but they suffer from some serious shortcomings. They all depend on ad hoc rules, and are capable of making serious mistakes by missing formant mergers or using spurious peaks. In addition, since linear prediction models only the vocal tract, nasalized sounds cannot be handled reliably.

Hidden Markov Models (HMMs) have been applied recently to the problem of formant tracking by Kopec (1986). He has used HMMs in which the states correspond to formant frequencies and the observations are LPC codebook symbols. The models are trained on hand-marked data from LPC spectrograms. The training process generates the transition and observation probabilities, which serve to effectively impose continuity constraints on formant movements. Kopec claims his model to be about 90% accurate, and supports this with detailed experiments. Once nice feature of his method is that he is able to make tradeoffs between spurious formants and missed formants by varying a single threshold value. In addition, there is no need for ad hoc rules. On the negative side, considerable effort must go into training the model.

Another recently reported method, due to Niederjohn and Lahat (1986), uses a bank of bandpass filters. A first estimate of the formants is made based on the energy output of the filters. This is followed by a statistical analysis of the consistency of the intervals between successive zero-crossings at the output of each filter. A decision function is applied to the data obtained to yield the final formant frequencies. The interesting feature of this technique is that it is designed to work in noisy environments, where techniques such as linear prediction have significant problems.

A number of auditory models have been proposed to enhance key features such as formants. Seneff (1985) has incorporated a synchrony detection measure that yields a pseudo-spectrogram in which the formants are more sharply defined. She suggests a gradient approach to tracking formants, where the upper and lower edges of the peaks are followed. This may prove to be a more robust technique and may avoid some catastrophic mistakes. Another perceptually motivated approach has been reported by Hermansky et al. (1985). In this method an auditory spectrum is produced by critical-band filtering followed by equal loudness curve preemphasis and intensity-loudness conversion. A low-order all-pole model is then used to extract F1 and F2'.

We propose to combine several of these ideas into an expert system for formant tracking. LPC analysis will be used for a first pass, possibly preceded by some perceptually motivated processing. This should give good results for voiced, non-nasalized regions. The next step is to determine the trouble areas. Unvoiced regions can be found by a simple ratio of high frequency energy to low frequency energy. Nasalized regions are more difficult to determine, but there are a number of features (Glass, 1984) which are characteristic. In addition, the results of the LPC analysis are likely to be suspect in areas where the peaks do not form continuous paths. To determine formants in nasalized regions, we plan to attempt a limited use of HMMs. A rule-based system will be used to coordinate all of these results to yield the final formant trajectories.

7.2.5 Pitch Tracking

Voice pitch, or fundamental frequency, can provide important information that is relevant to phonetic recognition, speaker normalization, syntactic class, and emphatic stress. Like formant tracking, pitch tracking has proven to be a surprisingly difficult problem. The most common pitch-tracking error is pitch doubling -- identifying the second harmonic instead of the fundamental frequency. This is a very serious error since the pitch value is off by a factor of two when this confusion occurs. The problem can not be remedied easily either by smoothing or by applying pitch-continuity constraints because the conditions that create the problem generally extend over several speech frames. Pitch doubling occurs primarily because of the influence of the first formant, especially for vocalic segments with low-frequency first formants. When the first formant comes close in frequency to the second harmonic, this harmonic can become strongly reinforced, creating a strong periodicity in the output signal at a rate that is twice as high as the fundamental frequency.

We have implemented a pitch-tracking algorithm based on the technique described by Markel and Gray (1976) that attempts to reduce the influence of the first formant through the use of inverse filtering techniques. The SIFT (Simple Inverse Filter Tracker) algorithm is a hybrid time-domain, frequency-domain approach to the pitch extraction problem. The basic idea of SIFT is to process a waveform in such a way that it will begin to approximate the glottal source waveform; that is, the speech waveform unmodified by the vocal tract resonances. A waveform of this sort has an approximately triangular shape. Given a waveform of this shape, the period can be detected by autocorrelating the waveform from 0 time lag out to approximately 17 msec time lag. The resulting autocorrelation function will show the normal peak at 0 time lag, and another prominent peak at the time lag of the basic period of a voiced sound. The criterion for existence of a peak in SIFT is variable, being higher at short time lags (high F0), and somewhat weaker at longer lags (lower F0).

The approximation to a glottal waveform is achieved in the following way. A 31 msec segment of the original waveform is taken. This segment is digitally low-pass filtered at approximately 1100 Hz, and is then down-sampled to 2.5 kHz by taking every fifth point in the segment. This is done to save time in the processing, since the higher-frequency components of the signal are not of importance in this form of pitch extraction. The resulting filtered and down-sampled segment is then Hamming-windowed, and the autocorrelation technique of linear predictive coding is used to design an inverse filter with 4 coefficients. These coefficients are used to filter the down-sampled segment. This filtering is performed on the unwindowed segment. The effect of these manipulations is to remove (or reduce) the effects of vocal tract resonances on the waveform. The resulting segment is then Hamming-windowed and its autocorrelation function is computed as described above. If a peak in the range 2.5 msec lag (F0 = 400 Hz) out to 16.5 msec (F0 = 60 Hz) is detected, using a variable threshold, it is considered to be a candidate for a pitch period. Some additional logic is then used to reject spurious peaks. SIFT reports a pitch value every 7.8 msec using a 31 msec analysis window, resulting in a 4:1 overlap.

The rms amplitude in decibels is also computed and saved, as well as the zero-crossing count. Voiced/voiceless decision-making is based on rms amplitude, the zero-crossing count, and the shape of the autocorrelation function. Typical voiced sounds average about 6-10 zero crossings per 8 msec frame, while unvoiced sounds usually show about 30-40 zero crossings per interval. By default, SIFT will set the pitch of an interval to zero if the number of zero-crossings is greater than 25, and/or the amplitude is less than 30 dB. These default values can be changed by the user with switches on the command line.

7.3 Automatic Phonetic Analysis

7.3.1 Background and General Approach

Despite the availability of powerful, high-speed digital signal processing techniques, researchers have found automatic phonetic analysis to be a very difficult problem. There is fairly good agreement on the characteristics of the speech signal that make automatic phonetic analysis a difficult problem. A thorough understanding of these problems is essential, since the nature of these complexities will impose very important constraints on the design of a phonetic recognition system. Described briefly below are five problems that must be dealt with in any speech recognition algorithm that intends to deal with continuous speech from multiple speakers. The description is based on a discussion by Klatt (1980).

1. Acoustic-phonetic invariance: Because of the phenomenon of coarticulation, individual speech sounds are often very strongly influenced by neighboring speech sounds. These effects are often very large and would seem to rule the most straightforward template-matching approaches to detecting phonetic segments.

2. Segmentation: Mental representations of words consist of sequences of discrete phoneme-sized units. However, spectrograms typically do not show evidence of discrete sequences of speech-sound types. While some speech sounds such as fricatives and stops show relatively clear acoustic landmarks, others speech sounds, such as sequences of glides, semivowels, diphthongs and nasals are extremely difficult to segment since the speech-sound types have a strong tendency to merge into one another. This makes it very difficult to locate boundaries between segments, which complicates the process of assigning phonetic labels to the signal. This is an especially difficult problem for "continuous speech" systems; that is, systems that do not require the speaker to pause between individual words.

3. Time Normalization: The durations associated with individual speech sounds are highly variable, being influenced by factors such as: (1) overall speaking rate, (2) word- or sentence-level stress, (3) locations of syntactic boundaries and (4) the phonetic characteristics of adjacent speech sounds.

This variability in segment duration complicates the process of pattern matching. Methods must be found which either ignore duration entirely, or allow comparisons to be made between segments of different duration.

4. Talker Normalization: Individual talkers differ from one another in a variety of ways, including differences in (1) the length and shape of the vocal tract, (2) voice pitch and a wide variety of other characteristics associated with the laryngeal source, (3) strategies for implementing particular sequences of coarticulatory movements and (4) a wide range of variation associated with dialect differences. These variations seem to cause very little difficulty for listeners, but very little is known about the talker normalization process in humans.

5. Phonological Recoding: This refers to the application of phonological rules that cross word boundaries. For example, the final /s/ of the word "this" would generally be produced as a palatal fricative (/sh/) in a phrase such as "this shoe." English includes a very large number of rules of this type, many of which cross word boundaries. These kinds of phenomena will cause obvious problems when the phonetic sequence is checked against entries in the dictionary. The problem cannot be solved simply by adding another entry to the dictionary because, in the general case, one would not want to accept "thish" for "this."

7.3.2 Template Matching versus Feature-Based Approaches

In general, there are two different ways in which the front-end of speech-recognition systems might be handled. All speech-recognition systems involve pattern matching of some kind. Some representation of the units that you intend to recognize -- words, phonemes, syllables, diphones, etc. -- have to be stored in memory. Methods then have to be developed for measuring the quality of the match between an unknown input utterance and the templates that have been stored.

One of the most important questions that must be answered in designing an automatic phonetic analysis system is to determine how the units will be represented. Very broadly, there are two choices. The first choice is to use the raw, unanalyzed spectrum -- for example a linear predictive coding (LPC) or Fast Fourier Transform (FFT) spectrum. Approaches that makes use of all of the details in the spectrum are called "template matching" or "low-level pattern matching." The second approach is called 'feature extraction,' and the basic idea is that, instead of using the whole spectrum, decisions are made about what aspects of the spectrum are most strongly associated with the phonetic characteristics of the utterances. For example, a feature-based system might make use of formant-frequency patterns to recognize resonant sounds such as vowels and semivowels.

The low-level template-matching scheme is the approach that is used in single word, talker-dependent systems. The speaker trains the system by producing each word in the vocabulary. The word templates are stored on disk as a sequence of spectra -- usually a vector of LPC coefficients that are sampled every 10 or 20 ms. When unknown words are spoken, the system makes calculations of the spectral distance between the unknown word and each word in the template dictionary. The same kind of pattern-matching approach could be applied to any size unit-- phonemes, syllables, diphones, etc. The advantages of this approach are:

1. Low-level pattern-matching techniques are computationally simple.

2. A very powerful method called "time warping" has been developed to handle the time variability problem.

3. It has been shown to work quite well for relatively small vocabularies and single speakers.

4. These methods do not try to make early decisions about what is and is not important in the spectrum. They preserve all of the spectral details and are therefore less likely to make low-level errors that would be propogated upward to higher-level modules in the system.

5. These methods do not require detailed knowledge in a variety of areas in which our scientific understanding is incomplete. For example, the feature-based approach will require the system designer, in one way or another, to make explicit decisions about what features or 'attributes' separate [r] from [l], [b] from [p], nasal consonants from semivowels. In some cases we will have incomplete knowledge of what these features are, and in other cases the contrast may be controlled by information that is very difficult to extract automatically.

But there are several drawbacks with this approach:

1. There is a general feeling that this approach is best suited to talker-dependent systems. This is because low-level approaches preserve virtually all of the details of the talker's speech rather than trying to extract just those aspects that are most relevant to the phonetic content of the speech.

2. There is a general feeling that low-level pattern-matching is best suited to small vocabularies. The reason is the same: These approaches preserve spectral detail. If steps are not taken to reduce data, these methods can become far too inefficient with large vocabularies.

3. Low-level pattern-matching approaches fail to make use of a good deal of information that is known about speech perception -- for example, the importance of formant frequencies in vowel

perception.

Although we have done some experimentation with a pattern-matching approach, the largest share of our work on automatic phonetic analysis has involved a feature-based approach. Despite the potential problems with feature-based analysis, it is our feeling that this approach is more likely to succeed for a system designed to handle continuous speech, multiple speakers, and a large vocabulary.

7.3.3 Phonetic Recognition Using Multivariate Statistical Methods

A feature-based approach to automatic phonetic analysis requires solutions to three kinds of decision-making problems. It is first necessary to determine how to segment the signal into units for later classification. It is then necessary to determine what acoustic characteristics can be used to classify the acoustic segments into phonetic categories -- i.e., what information can be used to separate a /d/ from a /t/, a /w/ from an /r/, or an /s/ from a /z/. Once it is determined what acoustic information should be used to make a particular category decision, it is then necessary to decide where to set boundaries between phonetic categories or, more generally, to assign scores to particular acoustic segments that reflect the relative probabilities that the acoustic segment is associated with a given phonetic segment.

Although this may seem at first to be the simplest of the three problems, it has not turned out to be a trivial issue. The decision-making procedure should meet at least two requirements. First, it should place the boundary between two categories in a location that minimizes categorization errors. Second, since classification errors are inevitable, the algorithm should preserve 'graded' information about the degree of category membership, or the probability that a particular set of measurements would be obtained for a signal from a particular category. For example, instead of simply setting an absolute threshold between /d/ and /t/, it would be very helpful to have an algorithm that reported that "the probability of /d/ is 80% and the probability of /t/ is 20%." A method that we are currently implementing attempts to meet these two requirements using a multivariate distance measure (MVD).

A simple example will illustrate the basic approach that we are using. Figure 7-15 shows measured values of voice-onset time (VOT) measured from instances of /d/ and /t/ (VOT is the interval from consonant release to the onset of voicing). VOT values for /t/ (and all other voiceless stops) are generally longer than VOT values for /d/ (and all other voiced stops). The simplest way to separate the two categories would be to set a threshold at the arithmetic mean of the two category means. A graded measure of the degree of category membership would be the linear distance of the measured value from each of the category means. However, this approach fails to make use of very important information about the relative variances of the two measurement distributions. For

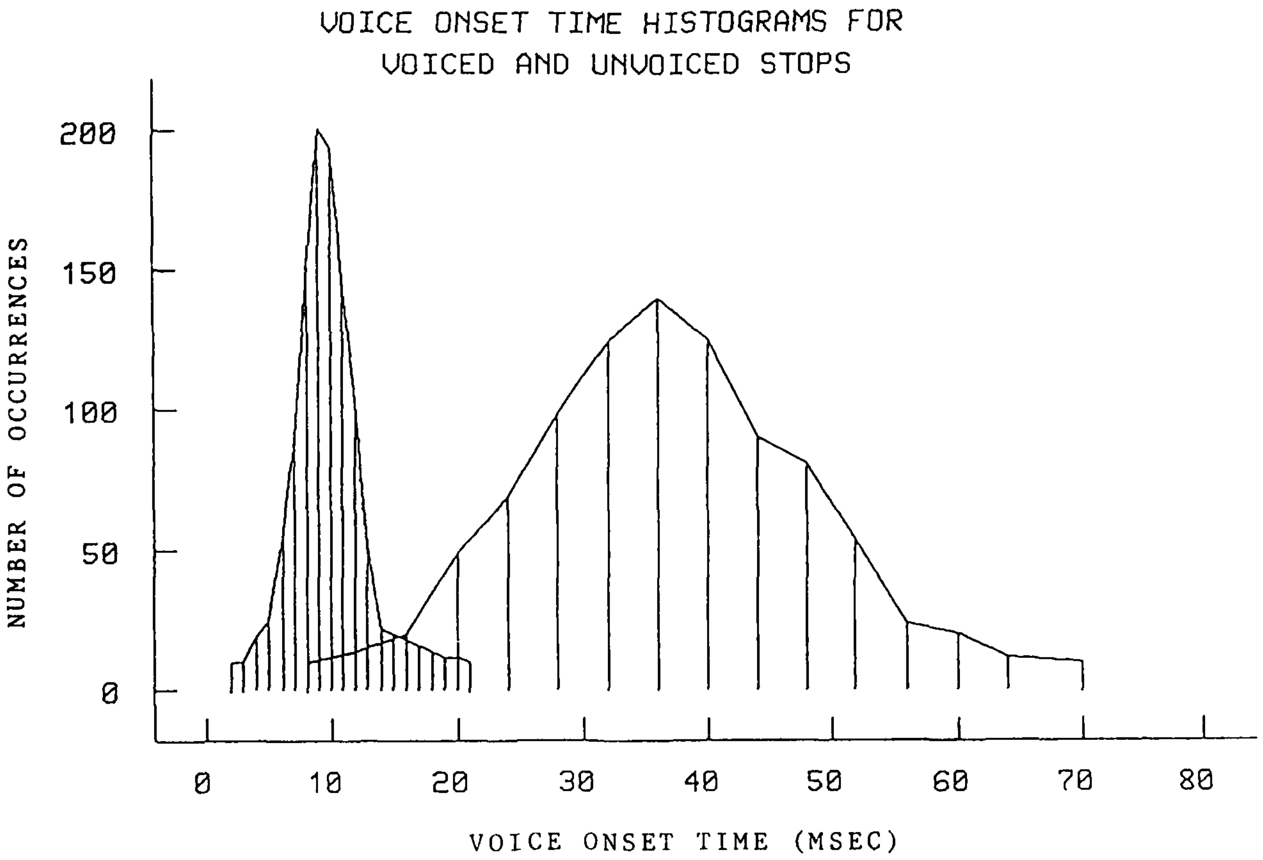


Figure 7-15. Distribution of voice-onset time values for voiced and unvoiced stop consonants.

example, because of the unequal variances, the optimal location of the threshold between these two categories would be the cross-over point in the two probability-density distributions rather than the mean of the two category means. Further, a more accurate measure of the degree of category membership would be the number of standard deviation units (i.e., z-score units) from each of the category means. In addition to minimizing classification errors, this method has the virtue of providing distance measures (z-scores) that can be interpreted directly in terms of statistical probabilities. Since this method uses a distance measure that is normalized in terms of relative variance, the technique can be used to compare distances from acoustic features with different measurement unit. For example, acoustic differences on a temporal feature such as voice-onset time can be either compared or combined with differences on a spectral feature, such as the onset frequency of the first formant, another feature that is known to be associated with differences in stop-consonant voicing.

The example shown in the figure represents the simplest case since the two categories are separated by measurements on a single acoustic dimension. There is excellent evidence in the speech-perception literature that human listeners make phonetic decisions by combining information from several acoustic dimensions. The pattern-classification method that we are implementing applies the same basic logic to the multivariate case; that is, where measurements are made on several acoustic dimensions. The multivariate case can be handled either by: (1) combining distance scores on several individual dimensions, or (2) calculating a single distance score in a multi-dimensional space. The first solution is computationally much simpler but, in theory, the second approach should be more accurate, since it takes cross-correlations among individual measures into account. The method that we have developed uses the second approach: a single multivariate distance measure is calculated that takes the inter-parameter covariance matrix into account.

The measures that are calculated represent distances between an unknown token (a point in an n-dimensional space) and the centers of any number of underlying phonetic categories (n-dimensional "elipsoids" in the n-dimensional space). Phonetic recognition is accomplished simply by selecting the phonetic category that produces the smallest distance with the respect to the unknown token. The phonetic categories are defined by training the system on acoustic feature values for known phonetic segments from a relatively large data base of diverse talkers. The distances that are calculated can be reported either as multivariate distances in chi-square units, or as chi-square probabilities. The chi-square probabilities represent the probability that a set of measurements from an unknown token could have been drawn from a given phonetic category.

7.3.4 Speaker-Independent Vowel Classification

The research described below was designed to determine the feasibility of using MVD in a speaker-independent phonetic recognition task. Our first goal was to attempt to recognize isolated vowels using voice pitch formant-frequency measurements from a large data base consisting of 29 adult male speakers and 27 adult female speakers. The research was designed to answer two questions: (1) can MVD be used to used to recognize vowels from a large group of talkers with minimal information about the individual talker, and (2) can MVD be used to determine the combination of acoustic features that results in the most accurate vowel-recognition performance.

The data base that was used in the recognition tests consisted of hand-measured acoustic parameters from 29 adult males and 27 adult females collected at Bell Laboratories in a classic study by Peterson and Barney (1952). The measures consisted of voice fundamental frequency (f0) and the frequencies of the first three formants (F1-F3). Measurements were made from two repetitions of each of ten vowels in the environment "h-vowel-d" (e.g., "heed, hid, head, had, hod, hawed, hood, who'd, hud, heard"). The acoustic measurements were made by hand from amplitude sections on a sound spectrograph.

Fundamental frequency and formant frequency measurements from the vowel [ae] are shown in Table 7-1. Also shown are four derived measurements that provide information about the average acoustic characteristics for a particular talker. These measures consist of mean fundamental frequency (mf0), and mean values of F1-F3 (mF1, mF2, mF3). The averages represent mean values for a particular acoustic measurement summed across both repetitions of each of the ten vowels for a given talker.

Examination of the data in Table 7-1 shows the substantial variability in absolute formant frequency values that is seen across individual talkers. For example, the range of F1 values for [ae] exceeds one octave (minimum = 514 Hz, maximum = 1110 Hz, range = 596 Hz) and the range of F2 values is just under an octave (minimum = 1470, maximum = 2560, range = 1090). Absolute formant frequencies are quite variable even within a category of talkers. For example, F1 values for [ae] within the group of female talkers vary from 650 Hz to 1110 Hz, a range equal to about 3/4 of a octave.

The formant frequency variability that is seen in Table 7-1 is due to inter-talker differences in the overall length and specific anatomical configuration of the vocal tract. Although the relationships among formants are more stable than absolute formant frequencies, normalization schemes based on formant ratios are only moderately successful. Further, Fant (1973) has shown that the scaling factors that relate the formant frequencies of one group of talkers to another (e.g., adult males versus adult females) are

Table 7-1. Measures of fundamental frequency (f0) and formant frequency (F1-F3) for the vowel [ae] for each of 29 adult male talkers and 27 adult female talkers from the Peterson and Barney (1952) data base. Also shown are average fundamental frequency (mf0) and formant frequency values (mF1-mF3) for each talker. The designation "M1-1" denotes male talker 1, first repetition; "M1-2" indicates the second repetition for that talker, etc. Dashes designate missing data.

Subject	f0	F1	F2	F3	mf0	mF1	mF2	mF3
M1-1	93	630	1710	2400	106	498	1418	2329
M1-2	94	658	1755	2305	106	498	1418	2329
M2-1	100	630	1770	2350	109	486	1438	2256
M2-2	105	630	1642	2170	109	486	1438	2256
M3-1	128	690	1610	2560	139	491	1288	2508
M3-2	131	700	1690	2580	139	491	1288	2508
M4-1	133	620	1710	2110	123	463	1306	2389
M4-2	124	660	1800	2150	123	463	1306	2389
M5-1	132	740	1810	2970	137	489	1576	2422
M5-2	145	630	1750	2480	137	489	1576	2422
M6-1	143	830	1720	2180	135	538	1455	2191
M6-2	135	810	1670	2300	135	538	1455	2191
M7-1	120	680	1470	2280	122	481	1314	2311
M7-2	119	620	1580	2320	122	481	1314	2311
M8-1	133	680	1958	2542	146	550	1619	2564
M8-2	141	708	1840	2535	146	550	1619	2564
M9-1	125	688	1600	2300	130	462	1373	2324
M9-2	122	660	1570	2380	130	462	1373	2324
M10-1	112	697	1610	2540	119	535	1451	2445
M10-2	114	684	1634	2510	119	535	1451	2445
M11-1	140	560	1820	2660	165	481	1470	2401
M11-2	180	580	1670	2540	165	481	1470	2401
M12-1	121	550	1570	2600	126	423	1386	2634
M12-2	120	530	1610	2650	126	423	1386	2634
M13-1	114	628	1837	2570	121	491	1437	2484
M13-2	111	622	1890	2560	121	491	1437	2484
M14-1	143	740	1800	2450	166	534	1445	2490
M14-2	162	775	1810	2200	166	534	1445	2490
M15-1	119	676	1670	2540	122	541	1456	2526
M15-2	125	725	1687	2500	122	541	1456	2526
M16-1	143	600	2000	2570	161	504	1453	2328
M16-2	138	590	1950	2460	161	504	1453	2328
M17-1	107	514	2140	2600	106	466	1549	2449
M17-2	106	552	1800	2500	106	466	1549	2449
M18-1	110	660	1650	2500	126	506	1406	2444
M18-2	120	624	1700	2475	126	506	1406	2444
M19-1	131	680	1685	2620	135	503	1423	2743

Table 7-1, continued

M19-2	133	680	1705	2490	135	503	1423	2743
M20-1	145	725	1700	2425	143	518	1394	2418
M20-2	127	710	1650	2220	143	518	1394	2418
M21-1	111	660	1600	2400	123	519	1418	2329
M21-2	120	720	1680	2430	123	519	1418	2329
M22-1	103	721	1680	2400	110	470	1384	2414
M22-2	109	750	1710	2440	110	470	1384	2414
M23-1	133	640	1773	2490	145	534	1482	2461
M23-2	133	640	1840	2560	145	534	1482	2461
M24-1	100	670	1860	2500	109	490	1433	2381
M24-2	100	670	1860	2500	109	490	1433	2381
M25-1	147	618	1735	2425	145	518	1429	2351
M25-2	123	615	1810	2400	145	518	1429	2351
M26-1	125	650	1738	2400	133	520	1528	2321
M26-2	130	663	1820	2400	133	520	1528	2321
M27-1	116	640	1620	2200	126	435	1326	2272
M27-2	118	650	1580	2360	126	435	1326	2272
M28-1	125	750	1610	2340	143	549	1437	2386
M28-2	136	770	1580	2350	143	549	1437	2386
M29-1	116	640	1710	2450	125	507	1444	2483
M29-2	128	592	1734	2480	125	507	1444	2483
F1-1	225	1040	1960	2920	227	638	1632	2947
F1-2	220	1010	1980	3080	227	638	1632	2947
F2-1	243	950	1970	2890	254	629	1661	2759
F2-2	244	980	1950	2920	254	629	1661	2759
F3-1	233	700	2560	3150	240	602	1818	2843
F3-2	225	675	2510	3145	240	602	1818	2843
F4-1	171	806	1970	2600	209	558	1739	2713
F4-2	150	825	1860	2550	209	558	1739	2713
F5-1	205	823	2220	2870	222	562	1792	2886
F5-2	200	800	2100	2900	222	562	1792	2886
F6-1	222	1110	2160	2700	201	601	1624	2764
F6-2	214	1070	1920	2750	201	601	1624	2764
F7-1	171	773	2000	2870	187	548	1726	2883
F7-2	175	875	2100	2970	187	548	1726	2883
F8-1	230	690	2185	2990	231	555	1764	2813
F8-2	220	660	2200	3020	231	555	1764	2813
F9-1	167	790	2180	3020	249	605	1784	2950
F9-2	280	840	2160	3020	249	605	1784	2950
F10-1	237	1020	1900	2960	251	613	1661	2804
F10-2	233	1005	2050	2870	251	613	1661	2804
F11-1	192	845	1700	2300	204	560	1552	2459
F11-2	187	860	1724	2530	204	560	1552	2459
F12-1	206	1008	1990	2870	231	645	1727	2864
F12-2	200	1040	2000	2800	231	645	1727	2864
F13-1	210	1010	2060	2900	223	581	1686	2740
F13-2	200	980	2160	2920	223	581	1686	2740
F14-1	188	750	2060	2770	184	532	1713	2664
F14-2	162	650	2110	2618	184	532	1713	2664
F15-1	236	873	2400	3060	252	590	1742	2963
F15-2	264	845	2380	3060	252	590	1742	2963
F16-1	187	940	2250	2760	206	559	1786	2779
F16-2	200	820	2200	2920	206	559	1786	2779

Table 7-1, continued

F17-1	192	860	1920	2850	210	571	1614	2922
F17-2	200	800	1980	2810	210	571	1614	2922
F18-1	233	860	2070	2880	250	571	1685	2848
F18-2	240	890	1920	2710	250	571	1685	2848
F19-1	224	784	1800	2750	231	572	1714	2721
F19-2	234	820	1750	2890	231	572	1714	2721
F20-1	218	808	2070	2880	225	570	1710	2765
F20-2	203	678	2420	3080	225	570	1710	2765
F21-1	189	850	1853	2685	202	579	1661	2582
F21-2	193	830	1800	2620	202	579	1661	2582
F22-1	205	900	2090	3000	214	589	1783	2837
F22-2	200	860	2160	2870	214	589	1783	2837
F23-1	225	1020	2030	2700	229	640	1776	2816
F23-2	225	1000	2200	2770	229	640	1776	2816
F24-1	212	710	2120	2600	212	551	1693	2541
F24-2	210	690	2250	2680	212	551	1693	2541
F25-1	194	810	1860	2620	246	556	1554	2678
F25-2	234	890	1800	2700	246	556	1554	2678
F26-1	200	960	2100	3000	264	631	1798	3004
F26-2	217	822	2200	3260	264	631	1798	3004
F27-1	187	861	2100	2800	216	525	1614	2796
F27-2	224	896	2040	3000	216	525	1614	2796

non-uniform and tend to show sizeable variations from one vowel category to another. Fant's findings indicate that the vowel normalization problem is not analogous to the relatively simple problem of transposing melody into a different key. This would seem to rule out normalization schemes based on the application of uniform scaling constants.

The recognition test using MVD and the Peterson and Barney data base was designed to address two questions:

1. Can MVD be used to classify vowels from a large group of male and female talkers?

2. Can MVD be used to determine what sets of acoustic parameters and parameter representations should be used to recognize vowels across talkers?

Using MVD involves training the program on each phonetic category to be recognized (the ten vowels in the present case) and then measuring distances between unknown tokens and each of the phonetic categories. For each unknown token, MVD reports both a distance and a chi-square probability to all ten vowel categories. The token is assigned to the category with the smallest distance (or highest probability).

Both the training and testing phases can make use of the entire data base, or any subset of the data base (e.g., only male talkers, or a random half of the talkers). Further, the system can be trained on all of the acoustic parameters in the data base, or any subset of the parameters. The second of the two questions listed above is addressed by determining which combination of parameters produces the best recognition performance.

7.3.5 Results of the Recognition Tests

Results of the recognition tests are shown in Tables 7-2 and 7-3. The data in Table 7-2 represent recognition accuracies using exclusively internal information; i.e., information that describes the characteristics of a particular token (i.e., some combination of fundamental and formant frequencies) without reference to any information that describes the characteristics of the individual speaker. The data in Table 7-3 represent recognition accuracies based on parameter sets that make use of both internal and external information, where external information consists of a measurement or set of measurements that describe the characteristics of an individual speaker. The external information is intended to normalize for differences across talkers. Examples include average fundamental frequency and average formant frequencies. In an actual recognition system, the normalization information would be gathered in a brief pre-enrollment session in which the speaker would be asked to produce a small number of standard utterances. Alternatively, the system could be designed to operate initially in a fully speaker-independent, and could then adapt to the individual

Table 7-2. Results of recognition tests using a multivariate distance measure (MVD) and acoustic measurements of ten vowels from the Peterson and Barney (1952) data base. MVD was either trained on all of the talkers in the data base (29 adult males and 27 adult females) and tested on the same set of talkers, or it was trained on a random half of the takers (i.e., the odd-numbered talkers) and tested on the other half. The acoustic paramter sets represent various combinations of voice fundamental frequency (f0), the frequencies of the three lowest formant (F1-F3), formant ratios (F1/F2, F1/F3) and log spectral distances (e.g., log F2 - log F1). All of the results shown in the table represent recognition based exclusively on internal information; i.e., parameter sets that describe the characteristics of the unknown token without reference to any information that describes characteristics of the individual speaker (e.g., average pitch, average formant frequencies, etc.).

Training Set	Testing Set	Percent Correct	Acoustic Parameters
all	all	81.2	F1, F2
odd	even	79.5	F1, F2
all	all	85.5	F1, F2, F3
odd	even	84.2	F1, F2, F3
all	all	89.5	f0, F1, F2
odd	even	88.6	f0, F1, F2
all	all	88.9	f0, F1, F2, F3
odd	even	88.5	f0, F1, F2, F3
all	all	77.5	F1/F2, F1/F3
all	all	69.9	logF1-logf0, logF2-logF1
all	all	80.7	logF1-logf0, logF2-logF1, logF3-logF2

Table 7-3. Results of recognition tests using a multivariate distance measure (MVD) and acoustic measurements of ten vowels from the Peterson and Barney (1952) data base. The data represent results for parameter sets that involved combinations of both internal and external information; that is, acoustic information describing the token and acoustic information describing the characteristics of the talker. The internal information consisted of various combinations of voice fundamental frequency (f0), the frequencies of the three lowest formant (F1-F3), formant ratios (F1/F2, F1/F3) and log spectral distances (e.g., log F2 - log F1). The external information consisted of average values for a particular talker, e.g., mean fundamental (mf0), or mean formant frequencies (mF1, mF2, mF3). For all of the results shown below, MVD was trained on all 54 talkers, and tested on the same set of talkers. Although not shown in the table, recognition accuracy was also tested for conditions in which MVD was trained on a random half of the talkers, and tested on the other half. (See text for details.)

Percent Correct	Acoustic Features
89.7	F1, F2, mf0
88.4	F1, F2, F3, mf0
91.7	F1, F2, mF1
91.2	F1, F2, mF2
88.3	F1, F2, mF3
94.4	F1, F2, mF1, mF2
94.5	F1, F2, mf0, mF1, mF2
93.9	F1, F2, mf0, mF1, mF2, mF3
93.5	F1, F2, F3, mF1, mF2, mF3
93.1	F1, F2, F3, mf0, mF1, mF2, mF3
93.8	F1, F2, mF1, mF2, mF3
82.7	F1/F2, F1/F3, mF1, mF2
82.3	logF1-logf0, logF2-logF1, mF1, mF2
92.1	*F1, F2, mF1-[aiu], mF2-[aiu]
93.5	**F1, F2, mF1-mid, mF2-mid
90.6	***F1, F2, F1-[i], F2-[i]

*Averages based on measures of the "point" vowels [a], [i] and [u].

**Averages based on measures of three central vowels.

***Normalizing information consisted of F1 and F2 of the vowel [i].

speaker after a sufficient amount of speech had been analyzed to allow estimation of average pitch and formant frequency values.

Table 7-2 shows the results for a variety of parameter sets that rely exclusively on the characteristics of the unknown token, without reference to any external speaker information. One issue that is addressed by the data in Table 7-2 concerns the degree of overlap between the training tokens and the test tokens. The entries labeled "all" in the table represent results for tests in which there was complete overlap between the training tokens and the unknown test tokens; the entries labeled "odd-even" represent results for tests in which there was no overlap between the training and test tokens. In every case, the system performs better when there is complete overlap, but the drop in performance is relatively small (0.4 - 1.7%, mean = 1.1%). These findings are encouraging since they indicate that system performance should remain good even when MVD is trained on one group of talkers and tested on another group.

Recognition accuracies for the 100% overlap entries vary from a minimum of about 70% to a maximum of just under 90%. Although it is well known that vowel perception by human listeners is very strongly associated with variations in F1 and F2, this simple parameter set yielded only moderately good performance (81.2%). The inclusion of F3 improved performance somewhat (85.5%), but it was primarily the addition of pitch information (f0) that resulted in recognition performance approaching 90%. In fact, when f0 was included, MVD performed slightly better without F3 (89.5%) than with F3 (88.9%).

The last three entries in Table 7-2 represent an attempt to represent vowels in terms of formant <u>relationships</u> rather than absolute formant frequencies, as has been suggested by a number of investigators. The first attempt tested Minifie's (1973) idea of representing the ratio of F1 to F2, and F1 to F3. The last two entries represent an attempt to test Miller's (1982) idea of representing distances between formants in logarithmic dimension. It can be seen that none of the approaches based on formant relationships worked as well as the straightforward parameter set consisting of absolute values of f0, F1 and F2. These results are somewhat surprising since: (1) schemes based on formant relationships are appealing on intuitive grounds, and (2) it has been shown that both of these approaches work well when tested on parameter values averaged across a given talker group.

Although not shown in Table 7-2, data were gathered to compare recognition accuracy on male versus female talkers. It was found that the system performed very similarly on the two groups. Averaged across all 11 conditions in Table 7-2, MVD yielded 82.5% correct on male talkers, and 83.7% on female talkers. Absolute male-female differences on individual parameter sets averaged 1.8%, and did not exceed 3.1%.

For all of the data shown in Table 7-2, male and female talkers were mixed, and no attempt was made to make use of information about the talker's sex. It was anticipated that MVD performance would improve if the system were trained and tested exclusively on male talkers, and then trained and tested exclusively on female talkers. This turned out not to be the case. For example, with the parameter set consisting of f0, F1, F2, MVD performance with in-class training/testing was identical to performance when male and female talkers were mixed. This finding is encouraging since it indicates that good performance does not require an apriori determination of the sex of the talker.

It is very clear from Table 7-2 that voice pitch significantly improves system performance. For example, recognition accuracy is about 81% with F1 and F2 alone, but improves to just under 90% with the inclusion of pitch information. There are two very different hypotheses regarding the importance of pitch information. According to Miller (1982), voice pitch is an integral part of the timbre of vowels, and therefore must be included in the parameter set describing this class of sounds. However, another possibility is that voice pitch serves primarily to provide indirect information about the acoustic characteristics of the speaker, and therefore functions primarily as a normalizing parameter. The reason is that voice pitch is strongly correlated with vocal-tract size. Therefore, when pitch information is included, MVD is essentially able to compare talkers with similar vocal-tract characteristics. If this is the case, it should be possible to find other acoustic characteristics, such as average formant frequencies, that are more strongly correlated with vocal-tract size, and might therefore function better as normalizing parameters. The data presented in Table 7-3 show the performance of MVD using a variety of different combinations of both internal and external information -- that is, combining information about the characteristics of the token and the characteristics of the talker.

It can be seen that there are a variety of feature combinations that produce recognition accuracies in the 90-95% range. It can also be seen that average F1 and F2 -- alone or in combination with one another -- produce better recognition accuracies than average fundamental frequency. A very simple parameter set consisting of F1, F2, mean F1 and mean F2 yielded 94.4% <1> correct recognition. The addition of average fundamental frequency yielded nearly identical performance (94.5%). It can also be seen in Table 7-3 that the schemes based on formant relationships (formant ratios and log formant distances) perform only moderately well even when normalizing is included.

<1> This figure is virtually identical to the 94.5% correct performanre of a panel of 70 human listeners who were asked to identify vowels from the same set in the original Peterson and Barney (1952) study.

For all except the last three entries in Table 7-3, mean frequency values for individual talkers were based on averaging all ten vowels in the data base. The last three entries in the table represent attempts to calculate normalizing information based on a smaller subset of the data base. For example, "mF1-[aiu]" indicates a measurement of mean formant frequency based on an averaging of F1 values from the vowels [a], [i] and [u] only (the vowels in "sod", "seed" and "sued"). This particular set was chosen because these vowels represent the articulatory and acoustic extremes of the English vowel space. It has been suggested that the so-called "point" vowels might be used by listeners to normalize for acoustic differences resulting from variation in vocal-tract size and configuration (e.g., Gerstman, 1968). It can be seen that MVD performs quite well (92.1%) using formant averages computed from these three vowels. However, performance is actually slightly better (93.5%) when averages are made from three centralized or "mid" vowels (the vowels in "bed, "bird" and "bud"). In fact, the last entry in the table indicates the recognition accuracy is quite good (90.6%) even when the normalizing information consists only of F1 and F2 of the vowel [i]. In general, the results from the last three entries in Table 7-3 suggest that the system should perform well even when normalizing information is obtained from very limited samples of speech.

7.3.6 Error Analysis: Preliminary Results

An important goal for future work with the MVD technique is to analyze the kinds of errors produced by the recognition system. It is clear from the results above that MVD is capable of relatively high recognition accuracies when information about the unknown token is combined with information about the acoustic characteristics of the talker. It is also clear, however, that recognition errors are inevitable with this or any other phonetic recognition system. For this reason, when the inevitable errors occur, it is highly desirable that the algorithm choose a phoneme that is phonetically similar to the correct phoneme. This is an important issue since error recovery at higher levels of the recognition system is a much simpler matter if incorrect phonetic segments are related in predictable ways to the correct segments.

Table 7-4 presents a confusion matrix -- a table showing how MVD distributed both correct and incorrect classifications -- for a parameter set consisting of F1, F2, and average F1. All values are in percent, and correct choices are shown along the diagonal. Analysis of the error patterns in Table 7-4 suggests that when MVD makes errors, it is very likely to choose a vowel that is phonetically similar to the correct vowel. For example: (1) when the input is "eh" ("bet") all of the errors are in the adjacent vowel categories "ih" ("bit") and "ae" (as in "bat"), (2) when the input is "uw" ("boot") all of the errors are in the adjacent category "uu" ("book"), and (3) when the input is "iy" ("beet"), all of the errors are in the adjacent vowel category "ih" ("bit"). These very logical error patterns should make it much simpler to

Table 7-4. Confusion matrix showing the distribution of correct and incorrect choices made by the MVD recognition algorithm using a parameter set consisting of F1, F2 and mean F1. All values are in percent; correct responses are shown along the main diagonal. The results indicate that when MVD makes errors, the incorrect vowel is always phonetically similar to the correct vowel.

		MVD Output									
		ae	ah	aw	eh	er	ih	iy	uh	uu	uw
Input to MVD	ae	92	0	0	8	0	0	0	0	0	0
	ah	0	90	5	0	0	0	0	5	0	0
	aw	0	4	95	0	0	0	0	1	0	0
	eh	4	0	0	89	0	7	0	0	0	0
	er	1	0	0	0	94	0	0	0	0	5
	ih	0	0	0	9	0	91	0	0	0	0
	iy	0	0	0	0	0	4	96	0	0	0
	uh	0	9	4	0	0	0	0	87	0	0
	uu	0	0	0	0	0	0	0	0	89	11
	uw	0	0	0	0	0	0	0	0	8	92

design a higher level module of the recognition system to recover from phonetic recognition errors.

7.3.7 Summary of MVD Recognition Tests

A multivariate statistical distance measure (MVD) was developed and tested on a large, multi-talker data base consisting of fundamental frequency and formant frequency measurements from two repetitions of each of ten English vowels produced by 29 male talkers and 27 female talkers. A number of derived measurements were also made which represented the acoustic characteristics of the particular talker (e.g., average fundamental frequency and average formant frequencies.) MVD was either trained on all of the talkers in the data base and tested on the same set of talkers, or trained on a random half of the talkers and tested on the other half. Further, the system could make use of all of the acoustic measurements, or any subset of the measurements. The results of the recognition tests included the following:

1. Recognition accuracies as high as 90% were obtained using exclusively internal information; i.e., acoustic measurements of the unknown token itself, without reference to any information describing the characteristics of the individual talker. The best internal parameter set consisted of f0, F1, and F2.

2. Recognition accuracies as high as 95% were obtained when internal information was combined with a very small amount of normalization information describing the characteristics of the individual talker. The most successful parameter set consisted of F1, F2, mean F1, and mean F2. However, there were a variety of other simple parameter sets that yielded recognition accuracies exceeding 90%.

3. The most accurate recognition was achieved when average values for individual talkers were based on an averaging of frequency measurements from all ten vowels. However, very good performance could be achieved when normalizing information was obtained from a much smaller set of speech samples. For example, 93.5% correct identification was achieved when formant averages were obtained from just three vowels, and 90.6% identification was achieved when formant averages were obtained from a single vowel.

4. A very small decrement in performance (1.1%) was observed when MVD was tested under conditions of no overlap between the talkers that were used to train the system and the talkers who were used to test the system.

5. Performance of MVD was virtually identical for male and female talkers. Further, it was shown that MVD did not need to be trained separately on male and female talkers.

6. Error analysis showed that when classification errors occurred, the incorrect token was always phonetically related to the correct token. These error patterns should facilitate the design of a higher level module of the recognition system to recover from phonetic classification errors.

7.3.8 Segmentation

As indicated in the introduction to the section on phonetic recognition, segmentation is one of the most difficult problems in automatic phonetic analysis. The reason is that individual speech sounds often merge into one another, meaning that clear acoustic boundaries between phonetic units arc only occasionally present. In general, there are two possible approaches to the segmentation problem that might be referred to as explicit and implicit segmentation. In explicit segmentation, the signal is first passed through an acoustic segmenter that hypothesizes starting and ending times for unlabeled phonetic units. These acoustic segments are then passed to other modules for labeling. Implicit segmentation does not involve an separate segmentation stage. Pattern-matching techniques are used to determine the quality of the match between a given portion of an utterance and a dictionary of templates. Segmentation decisions are made post-hoc based on the range of speech frames over which the utterance and template match. We are currently exploring both explicit and implicit approaches to segmentation. Two explicit segmentation schemes will described briefly below.

Figure 7-5a shows a segmentation scheme based on a very straightforward measure of the rate of spectral change. The program calculates the bin-for-bin difference between the current (smoothed) FFT spectrum and and another spectrum some number of frames downstream. A step size of 10 msec was used for the "spectral derivative" function shown in Figure 7-5a, based on the word "obey." It can be seen that the function shows peaks at the onsets of the two vowels, and a trough at the juncture between the first vowel and the consonant. The function is also relatively flat within the consonant and vocalic segments.

We have also experimented with a more complex segmentation scheme called the "association waveform," based on the work of Glass and Zue (1986). The algorithm attempts to "...associate a given frame with its immediate past or future..." (Glass and Zue, p. 26) using correlational techniques. Although the algorithm is much more complex, we have found that it produces functions that are very similar to the simple spectral derivative calculation described above.

If we choose to pursue one of these explicit segmentation schemes in favor of an implicit segmentation scheme, it will be necessary to develop methods for setting spectral-change thresholds for the detection of segment boundaries. This will be done by collecting a large body of statistics on the characteristics of spectral change within and across segment boundaries. Probabilities could then be assigned to spectral-change functions using the multivariate statistical techniques described above under the section on segment labeling.

7.3.9 Measurement of Phoneme Distance

Since the phonetic string generated by the acoustic-phonetic module is errorful, procedures for error recovery are essential. To a large extent, errors can be handled by designing an acoustic-phonetic that generates relative probabilities rather than segment labels. The lexical module can then hypothesize all phonetic labels whose probabilities exceed a certain threshold. However, in order to handle case of labeling errors in which the correct label's probability does not exceed this threshold, it is necessary to incorporate general measures of phoneme similarity.

The phoneme similarity measure that we have developed is based directly on measured perceptual similarities between phonemes. The matrix is based on perceptual confusion data reported by Miller and Nicely (1955), which were then submitted to multidimensional scaling analysis (Shepard, 1980). This matrix, shown in Table 7-5, correlates strongly with simpler phoneme distance measures that are based on the number of shared phonetic features. However, the matrix that is based directly on measured perceptual similarities is able to capture certain similarities that are not predicted by the feature approach (e.g., [b]-[v] and [p]-[f]).

7.3.10 Future Directions

Two relatively large-scale studies are planned as a follow-up to the multivariate classification work that was carried out with the Peterson and Barney (1952) data base. The two most significant findings from the work with the Peterson and Barney data base are: (1) the multivariate statistical approach to phonetic classification is capable of very high levels of recognition accuracy when tested on a large and diverse group of talkers, and (2) the technique offers a very powerful way to determine <u>empirically</u> what combinations of features produce the best recognition accuracies. However, there are two important limitations to these findings. First, the fundamental frequency and formant frequency data were measured by hand. This was crucial for the initial tests since the simplification enabled us to separate the problems of recognition and feature-set selection from the low-level, feature-extraction problems. However, future work will incorporate the automatic pitch and formant-tracking methods that we are developing, described in Section 7.2. The second limitation to the vowel-classification results is related to the fact that the acoustic-feature measures were essentially <u>static</u> -- that is, the measures represented a single spectral slice of the signal. Obviously, the recognition system will need the capability to analyze patterns whose feature values change over time. A project that is just underway is aimed at extending the MVD

Table 7-5. Phoneme-to-phoneme similarity matrix based on a multidimensional scaling analysis (Shepard, 1980) of Miller and Nicely's (1955) data on the perceptual confusions among English consonants. The numbers are in arbitrary units. ,(Note: th- = voiceless th; th+ = voiced th.)

	k	p	t	f	th-	s	sh	v	th+	z	zh	g	d	b	m	n
k																
p	10															
t	16	17														
f	34	24	38													
th-	32	22	33	10												
s	42	32	33	10												
sh	51	44	37	50	40	24										
v	69	60	69	38	38	39	60									
th+	81	71	80	50	49	46	65	13								
z	88	78	83	61	57	48	60	23	18							
zh	108	98	103	82	78	68	77	43	36	22						
g	102	93	100	72	71	66	80	35	22	21	23					
d	107	97	106	76	77	73	89	38	27	31	31	12				
b	77	68	80	43	47	53	75	17	22	40	57	38	37			
m	83	77	93	56	78	79	113	54	63	80	97	78	73	21		
n	100	93	109	70	65	90	103	58	62	80	92	71	64	42	21	
	k	p	t	f	th-	s	sh	v	th+	z	zh	g	d	b	m	n

technique to vowel classification in dynamic context; that is, utterances-in which parameters changes over time.

The plan is to measure the F1 and F2 formant frequencies every 10 to 20 msec for appropriate sample words. Two other parameters that will be used are the mean of F1 and the mean for F2 for each speaker. The mean, variance and covariance matrix will then be calculated for each point along the track. The words used will be of the consonant-vowel-consonant form, manually labeled and segmented. Initially, seven classes of vowels will be looked at, with each class containing approximately 60 samples of a vowel sound. Additional classes will be investigated that combine vowels with the consonants [l, m, n, r], since these combinations tend to cause more problems than others. This model should work well with classes of diphones. As with the MVD, the DMVD (Dynamic Multivariate Distance Measure) will produce both relative and absolute chi-square probabilities of the combinations.

One new problem which arises is the time alignment of these sampled tracks. Fortunately, there has been an important breakthrough in the past few years by the development of the Dynamic Time Warping technique. This process attempts to time align tracks by compressing portions of each track to reduce the overall sum of distances between the two segments. Time alignment has proven to be very successful in aligning tracks out of sequences by as much as 150 msec.

A second goal for the coming year will be to extend the MVD technique to other speech-sound classes. Although we have tested the technique only on vocalic sounds so far, the classification algorithm should work with all classes of speech sounds. Pilot data that we are now collecting represent an attempt to extend MVD to fricative detection and classification. The study will be an extension both of the MVD vowel-classification study described in this section, and of the expert-system approach to fricative classification described in Section 7.4. The expert system attempted to model human spectrogram reading abilities and identified a number of spectrographic features that used by spectrogram readers in classifying the place-of-articulation and voicing properties of fricatives. Left unresolved by that study were several issues related to: (1) calculation of some of the acoustic features that were found to be important for fricative recognition, (2) setting numerical thresholds between categories, and (3) normalizing for differences in phonetic context and talker. After developing the appropriate feature-extraction algorithms, we plan to use MVD to address the threshold and normalization problems. The overall strategy and design of the study will be similar to the speaker-independent vowel-classification work described in this section.

A third goal for the coming year will be to determine what level of formant-frequency measurement accuracy is needed attain relatively high levels of phonetic-recognition accuracy. It is quite clear that formant-tracking errors of a 5 or 10 Hz are unimportant, and that errors of several hundred Hz are almost certainly very important. However, as far as we are able to determine, no study has addressed the question of the relationship between formant-frequency measurement error and phonetic-recognition accuracy. With the Peterson and Barney data base and the MVD classification algorithm, we are in an excellent position to address this important issue. Our plan is to use a random-number generator to introduce specific amounts of error in the formant-frequency data, and then to measure the effects of these errors on phonetic-recognition accuracy.

7.4 Expert System for Phonetic Classification

7.4.1 Background

Human speech can be viewed as the conversion of muscular energy to acoustic energy. The muscular energy is used to induce a pressure change in the vocal tract and set the air into motion. It can then be used to regulate the flow of air or modify the sound waves (Catford, 1957).

After the airflow is generated, it passes through the laryngeal cavity, which consists of the larynx, vocal bands and glottis (Francis, 1958). Vibration of the vocal cords by the passing air, is known as voicing. The air then passes through the pharyngeal cavity where it can be routed through the oral or nasal cavities. Each of these can be vibrated and has its own resonating characteristics. In the oral cavity are the articulators, structures which break up or interrupt the air flow (Denes, 1975). The upper articulators consist of the upper lip, the upper teeth and the entire roof of the mouth including the alveolar ridge, the palate, and the velum. The lower articulators include the lower lip, lower teeth and tongue which is divided into areas called the tip, blade, front, dorsum and root (Catford, 1957). By employing these structures, the airflow can be altered to create a variety of sounds. Sounds have features that bear a direct relationship to the articulatory gesture which produced the sound.

Although the number of possible sounds is enormous, the actual number of basic sounds in a given language is quite restricted. In English there are about forty basic sounds, referred to as phonemes. Each phoneme has distinct properties according to the place of articulation, manner of articulation and voicing. These features serve to classify the phonemes.

Manner features are indicators of how the sound is made. A stop sound, as in the beginning of the word "to", is produced when the airflow is actually stopped. Pressure is built up and then released. A fricative is characterized by turbulence caused by a constriction in the airflow. The /f/ sound in "foo" is a fricative. Nasals are created when the nasal cavity is brought into play and can be heard in the consonant sound in the word "no".

Glides and liquids are sounds produced by forming some constriction in the vocal tract. The constrictions are smaller than those for vowels, but are still large enough so no turbulent noise is created. The sounds found in the beginning of "woe" and "yet" are typical of this class.

The place features of a sound refer to the position of articulatory mechanisms during the production of sounds. Labio-dental, alveolar, palatal, velar, dental, and palato-alveolar are terms used to describe the place features of the consonants, implying involvement of the lips and teeth, alveolar ridge, palate, velum, teeth, and palate and alveolar ridge respectively (Denes, 1975).

Many of the basic sounds have voiced-voiceless cognates, sounds that have the same place and manner features, but the vocal bands are vibrated when creating one of the cognates. Adding voicing to a voiceless sound creates a new sound. As an example, the /s/ sound is a voiceless fricative with /z/ as its voiced cognate.

Although phonemes have been studied for some time, the greatest advances have been made since the 1940's when the sound spectrograph was invented. This device makes visible records of the fundamental dimensions of speech: frequency, intensity and time. This record, called a spectrogram, is produced using frequency as the vertical axis and time as the horizontal axis. Variations in intensity are depicted by the darkness of the pattern.

7.4.2 Spectrogram Reading

A spectrogram displays spectral manifestations of physical speech producing actions (Figure 7-16). A stop is viewed on a spectrogram as a silent gap of about 20 to 150 msec, flow and is frequently followed by a burst of energy. A fricative, on the other hand, is generally signaled by a broadband noise lasting about 80 to 200 msec. Voicing is sometimes seen as a low-intensity band of low-frequency energy. Vowels, which are sounds made in a relatively unconstricted cavity, are indicated on a spectrogram by horizontal resonance bars called formants. Relationships between these formants are clues to the identity of the individual vowels.

Trying to determine the utterance recorded in a spectrogram is not a simple task. Speech produces a complex acoustic signal that contains extra-lingual material as well as the linguistic message. Speaker attributes such as physiology, sex, age, emotional state, and even whether the speaker is suffering from a cold, may be reflected in the speech signal. To complicate matters further, speech signals will differ not only from speaker to speaker but also in repetitions of the same utterance spoken by the same speaker (Hecker, 1971).

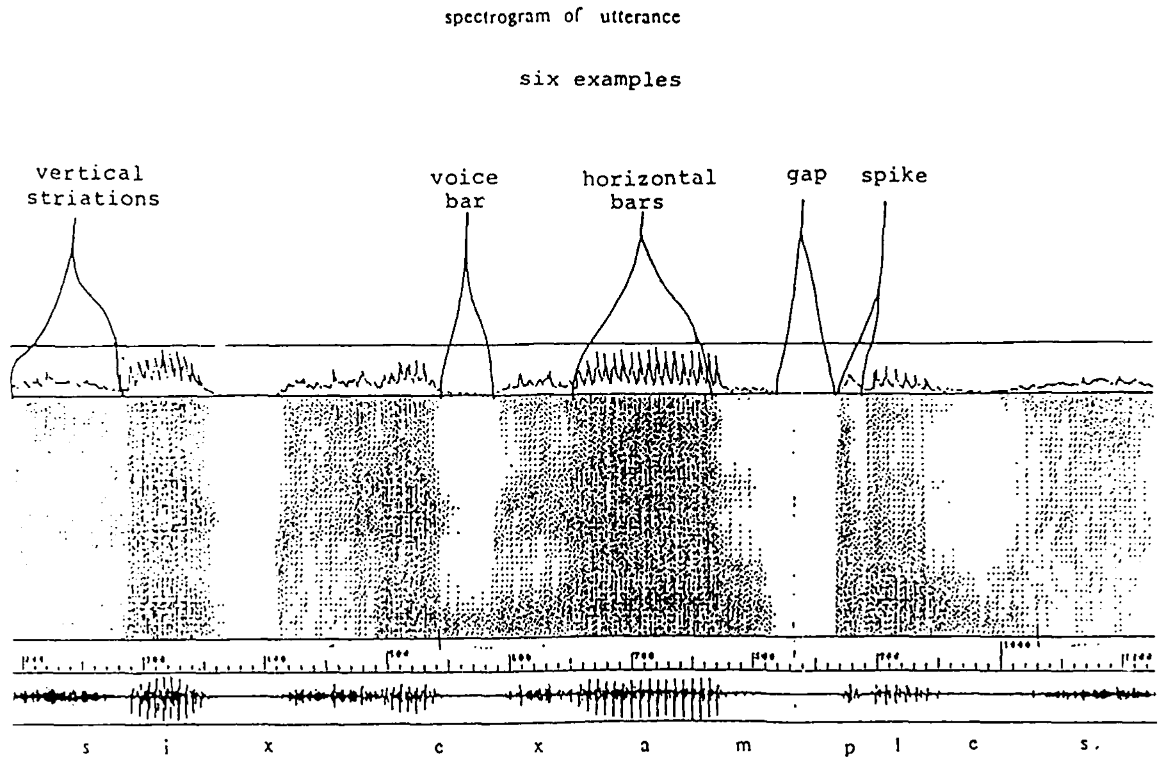


Figure 7-16. Spectrogram of the utterance "six examples" showing typical spectrographic landmarks used by spectrogram readers.

In addition, in continuous speech basic sounds are combined, causing a blurring of boundaries and properties of the individual sounds. This effect is referred to as coarticulation. In analyzing connected speech, one must be careful to distinguish coarticulatory effects from speaker variability (Hecker, 1971). Each phoneme has unique articulatory and acoustic properties which change with the phonetic environment. This overlap of phonetic information in the acoustic signal makes the spectrogram difficult to interpret and prevents a simple template matching solution (Zue, 1985b). When sounds are made in continual speech, the motions of the structures involved are continual. Templates made in disconnected speech are not truly representative of the sound as found in continual speech, as the patterns do not reflect coarticulatory effects.

To determine the phonemes in a spectrogram, boundaries that pinpoint changes in spectral composition must be determined. Next these segments should be labeled by classifications such as fricatives, stops and vowel-like sounds (Fant, 1957). Each of the segments is analyzed according to its classification. There are cues that help spectrogram readers identify the particular phoneme present in a segment. The information found in consonant segments tends to be more reliable than that found in vowel segments, as vowels suffer more from coarticulatory effects. Therefore it is usual to identify the consonants first and then use that information to help identify the vowels.

Many rules about phoneme spectral patterns are known, but to be used effectively the reader must know when to apply these rules and when to ignore them. Visual features may occur in spectrograms for a variety of reasons. The duration of a sound could be a clue to the identity of a phoneme, but could also indicate stress or the voicing feature of an adjacent sound. Some vowels are short by nature, like the schwa, while others can become shortened because of their environment. Vowels are known to be shorter when preceding a voiceless sound than when preceding a voiced sound (Potter, Kopp, and Green, 1966). Stressed vowels tend to be longer in duration than unstressed vowels, and if a final syllable is unstressed, its duration is quite short (Klatt, 1976). Also, certain consonant clusters significantly abbreviate or totally delete a member consonant. While a /t/ sound is classified as a stop, in the word "butter" the /t/ sound is usually pronounced as a fast stop as closure is not complete. The characteristics of the sound are significantly different from the stop /t/ and it is known as a flap.

With all the problems inherent in reading a spectrogram, the multitude of rules which must be selectively applied, there are those who have hypothesised that it is not possible to train a person to read them effectively (Klatt and Stevens, 1973). To settle this controversy, an experiment designed to assess a person's ability to read unknown utterances from spectrograms was conducted at Carnegie-Mellon University in late 1977 and continued in early 1978 (Cole, Rudnicky, Zue, and Reddy, 1980). Victor Zue, a leading researcher in the field of speech recognition who began a systematic study of spectrograms in 1971, was chosen as the subject who would attempt to identify utterances from their spectrograms.

Zue's task was to identify the phoneme strings represented by 23 spectrograms of utterances by two male speakers. First he identified segment boundaries and then labeled them phonemically. Three trained phoneticians also transcribed the utterances, but they did so by listening to them. As transcription is not an exact science, certain guidelines were agreed upon. A segment was said to exist when two of the three phoneticians agreed on its existence. When Zue was not sure of a phoneme label, he gave a second choice. His labels agreed with at least one of the phoneticians 85% of the time. As the average agreement among the phoneticians was only about 90%, Zue's performance was good.

Observation of this performance led to the conclusion that phonetic segments can be identified by characteristic visual patterns. Zue's extensive knowledge of the effects of coarticulation on these patterns was deemed instrumental in the interpretation of the spectrograms (Cole et al., 1980).

Further evidence exists in the presentation to the 1979 *International Conference on Acoustic Signal* and Speech Processing by Zue and Cole. They cited the results of a thirteen week course in acoustic phonetics that trained five spectrogram readers with a combined accuracy of 80%.

Their experiments demonstrated that phonemes are accompanied by acoustic features that are recognizable on a speech spectrogram, and that with sufficient training it is possible to learn enough about these features, and the modifications they undergo in fluent speech, to read a spectrogram of an unknown utterance (Zue and Cole, 1979).

Although phonemic recognition is a subjective task, the utterance can be identified even from the imperfect transcription. In the experiment using Victor Zue as a subject, the phoneme sequences were examined by a linguist who was able to identify all but 5 words from the fifteen utterances. Therefore using phoneme identification as a step in a speech recognition project provides a firm platform from which a linguist can do word recognition.

7.4.3 Expert System for Fricative Classification

The goal for this project was to build a phoneme recognizer using the data available from spectrograms and the methods used by spectrogram readers. Gathering information on the spectrogram-reading process is a difficult task. A few books, manuals and papers do exist on the methodology (Potter et al., 1966; Fant, 1957; Zue, 1985), but there is no exhaustive listing of the process. There is a large number of well-established rules concerning coarticulation and its effects on a visual representation of speech.

Two speech scientists in the Rochester area agreed to serve as informants for the expert system project. Robert Houde is a recognized authority in speech science and speech signal processing. He has been working in the domain of speech since 1957 and received his Ph.D. in communications science from the University of Michigan in 1967. Dr. Houde founded the Center for Communications Research in 1970, an organization researching concerns of the deaf. In 1983 he started Speech Recognition Systems, Inc. (SRS), and has been attempting to build a speaker independent continuous speech recognition system. Although Dr. Houde has been working with spectrograms for many years, he has not made a practice of reading them.

The services of Dr. James Hillenbrand were also available. Currently a research scientist for the RIT Research Corporation, Dr. Hillenbrand received his Ph.D. in Speech and Hearing Science in 1980 from the University of Washington. He has been involved in research in speech acoustics and speech perception since 1975.

Background research took a good amount of reading as a wealth of knowledge about speech is available. Once a basic understanding of the domain was achieved, a meeting took place with both experts. Knowing that the task of recognizing all the phonemes was well beyond the scope of a Master's thesis, it was decided, upon Dr. Houde's recommendation, to attempt recognition of the fricatives.

Although the fricative classification sometimes includes the /h/ sound, Zue's description of phoneme classifications state this sound is an aspirant and does not fit easily into any category (Zue, 1985). Knowing this and the fact that spectrogram readers usually detect /h/ sounds by the effect on formant transitions and not because of frication (Potter et al., 1966), the decision was made to omit /h/ from the fricative list.

The first version of Fric assumed that all fricatives would be positively identified and that the system would be interacting with an expert looking at a spectrogram. This totally interactive system guided the expert in doing feature extraction on the spectrogram. The design aim of Fric was to replace questions to the expert with automatic routines that could answer these questions by analyzing the information provided by SpeechTool (Section 7.2).

When the experts interacted with this version of Fric, a number of questions were raised about the intent of the program and the reasoning behind it. One problem that was identified at this point was that of ambiguous terminology. For instance, when asking about the amount of total energy in the sound segment, the system asked, "Is the sound weak or strong?". Weak and strong have phonetic connotations that caused the expert to think the system was asking about something other than how much energy was in the signal. It is possible to have a strong fricative with a small amount of energy.

Subsequent interviews were held with only one of the experts at a time. During these sessions the expert read spectrograms and was questioned about his reasoning techniques in an attempt to formulate the rules and control strategies needed for Fric. These sessions were taped to minimize loss of any information.

The next interview took place at Dr. Houde's place of business where he was recorded reading spectrograms for about an hour and a half. Knowledge about reading spectrograms was gained as well as guidelines about choosing sample utterances. It was necessary to separate similar utterances to prevent Dr. Houde from template matching. His work tends to use template matching as a fundamental concept and though Fric's strategy was to avoid that approach, Dr. Houde tended to use it whenever possible.

When reading spectrograms Dr. Houde would frequently use knowledge that would not be available to Fric. He knew the focus was on fricatives, and this slanted his answers. He also used his knowledge of the English language to fill in gaps in the transcription of the spectrogram. The following are quotes of the interview with Dr. Houde on 4/7/86.

> And there's a fricative here because there are fricatives everywhere.
>
> So maybe this says "if this", and this is all front vowel too, so this is strong so this could be "is". Maybe "if this is iii" and I could put a t on the end just because it would make linguistic sense. "if this is it", but I don't have any other reason to put the 't' on it. I don't think that is a 't'.

When the sample utterances were single words, Dr. Houde's ability to transcribe the utterance phonetically diminished. This can be attributed to the inability to use any higher level knowledge. It was decided that the next set of examples should be words strung together without making sense, so these higher level knowledge sources could not be used.

Dr. Houde's requests for subsequent sessions included a scale of 6 KHz on the spectrogram rather than the 4 KHz in use because many strong fricatives (/s/ or /z/) contain most of their energy above 4 KHz. Dr. Houde could discern the presence of these strong fricatives by the amount of total energy, which is shown in another graph, the sum function, even though there was no corresponding darkness on the spectrogram. Also requested was a print of the spectrogram made with the time scale set so individual pitch periods could be seen. When the compressed time scale made it impossible to detect periodicity, Dr. Houde was unable to tell if a sound was voiced.

This request indicated that although voicing may have many effects on a sound, one criterion, periodicity, was enough to discern voicing. During the background research many of the rules concerning voicing used its presence as a prerequisite condition leading to a particular consequence. What the system needed and now had was a rule which said if a certain acoustic phenomena is present, then one can conclude voicing is also present. Dr. Houde explained that he had a number of people working for over a month trying to develop a good algorithm to detect periodicity in the waveform, but was finally satisfied with the output. Consequently later efforts were concentrated on finding ways to discern the place of articulation.

Subsequent interviews required reading spectrograms that were designed to elicit specific information. The evolving set of rules was:

Discriminate between weak and strong fricatives.

If strong, look at the pattern to see where the energy is concentrated.

If weak, take a guess.

Look at waveform periodicity to detect voicing.

Dr. Houde could not discriminate among weak fricatives by any method other than template matching. If a weak fricative had been previously identified, he would compare its spectral composition with the segment under question. As template matching is not a desirable method in a speaker independent system and expert spectrogram readers have no need of templates, the next interview, which was held with Dr. Hillenbrand, attempted to find rules that would help to discriminate between weak fricatives.

From this interview came the idea that the spectrogram was really a reflection of the movements of the articulators. Dr. Hillenbrand is so aware of how structures move to create speech that he can follow that motion in the formant transitions. Watching the changes in formant directions was the key to distinguishing among the weak fricatives.

7.4.4 Rule Generation

All the knowledge gained from interviews needed to be embedded in a program. RuleMaster, a software tool for building expert systems, a product of Radian Corporation, was used to implement Fric. The RuleMaster expert system building package contains two principle components: Radial, an interpreted language for expressing and executing rules, and RuleMaker, which induces Radial rules from example tables.

RuleMaker removes the duty of rule generation from the shoulders of the system builder because, given specific examples, it can produce a rule to cover the situation (Reise and Zubrick, 1985). Intelligent editors are included to aid in the development of example tables and the special files needed to implement the hierarchical control of the expert system. "INDED" works with induction files that include the example tables and "SYSED" helps to manage the overall structure of the system.

Each module is provided with the ability to feed information to the explanation facility. This facility allows the user of the system to ask why certain information is being requested or why a conclusion was reached. The explanation supplied is dependent on the system structure and creator supplied intent statements. The suppression of certain information can be achieved if the system designer feels the information will confuse the user rather than explain the actions of the system.

The Radial interpreter can interface with external routines written in a variety of programming languages and with external information sources other than programs such as databases, instruments and other computers. This provides a powerful interface to existing machinery and programs that may already be in use. Fric is to interact with routines that can do feature extraction from a spectrogram.

The final version of Fric is a refinement of the first interactive system. This final Fric begins to answer some of its own questions. To determine the identity of a fricative phoneme, Fric first asks the user some basic information about the segment in question: the name of the file containing the utterance and the time boundaries delimiting the segment. Then a C routine is called that determines whether the amount of energy in the segment designates a strong or weak fricative. Inquiries are made about the spectral shape of strong segments, to be sure they are not weak segments that have a lot of energy. The threshold to determine weak or strong was set very low so no strong fricative would be misclassified as weak.

When a segment is classified as weak, Fric asks questions about the formant transitions to determine the place of articulation. The module dealing with segments labeled as strong first asks for limits that will act as boundaries for expected areas of concentrated energy and then uses a C routine to determine where the concentrations of energy are. These boundaries are needed as input as the areas of concentrations differ significantly for male and female voices.

Once the place of articulation is determined, information is requested about voicebars and periodicity in the waveform. This establishes the presence or absence of voicing. When the place of articulation is known and a determination on the presence of voicing has been made, enough information exists to uniquely identify a fricative.

7.4.5 Performance Testing

The amount of testing Fric was subjected to was limited by the memory requirements of the speech data. The data files require approximately 190 Kbytes for 1 second of speech. The system on which Fric was running had been operating at about 97% capacity for months; consequently few data files could be generated at one time. Testing was successful in indicating shortcomings of the system and showing this approach to be a valid method for phoneme identification.

The testing process involved obtaining speech data by having various speakers talk into a microphone in a relatively noise free environment room for test phrases. The utterances (Table 7-6) were well-articulated samples of continuous speech and some single words which were then processed by SpeechTool. Because the segmenter which will supply Fric with data has not been designed, it was unreasonable to attempt to duplicate non-fricative data which could be mistakenly sent to Fric. Therefore no data with a low zero-crossing rate, which is considered non-fricative, was used.

A total of 43 fricatives from 4 speakers, three male and one female, were identified by Fric. An identification was considered correct when Fric classified the sound as a single phoneme and the classification agreed with that of the tester.

Correct identification was made 60% of the time. Included as incorrect were cases where Fric did not have enough knowledge to decide between competing candidates and therefore gave two choices as to the identity of the phoneme. Of the identifications considered incorrect, 41% were classifications which gave correct information about the segment, but did not identify it as a single phoneme. It is reasonable to assume that with more knowledge gained from continuing the interview process, more of these general classifications could be narrowed down to a specific phoneme.

Of the incorrect identifications, 41% were strong fricatives that were misclassified as weak by the C routine. There are two possible remedies to this problem. One is to lower the energy threshold required to classify a segment as strong and the other is not to overlook the implications of the spectral shape when the segment is classified as weak.

Since spectrograms of male voices are known to be easier to read, it was reasonable to expect Fric to function better with male voices. Indeed, in utterances by male speakers, fricatives were correctly identified 74% of the time whereas with female speakers the correct identification rate was only 37%.

Table 7-6. Utterances used in the fricative-recognition tests.

speaker	fricatives	noisy sounds	identified	phrase
male	2	2	2	six
male	5	5	4	three free Sunday shows
male	2	1	1	five
female	4	4	2	fathom zoophyte
female	2	2	0	six
female	4	2	1	shove biff over
female	4	4	1	thin feathers
female	2	2	0	thither
female	2	2	2	five
male	4	4	3	six of these
male	3	1	1	shove biff
male	5	5	3	sue fixed the glass
male	5	5	3	the fifth oaf is
male	4	4	3	vote then fresh

Test Phrases

Most of the errors that were made with female voices were because of a poor initial diagnosis of the amount of energy in the speech signal. Apart from the threshold problem previously mentioned, this misdiagnosis could be due to the processing of the speech signal. SpeechTool accepts energy up to 6 KHz for processing, but many female voices have ranges up to 8 KHz. Since the strong fricatives (/s/ and /z/) have a concentration of energy in the high frequencies of the speaker's range, it is possible that the concentrations expected for strong fricatives from female speakers are being cut off by the processing of the data passed to Fric. With this amount of the speech signal missing, the sound is identified as weak. In order to remedy this problem, SpeechTool should work with frequencies up to 8KHz.

Even with these problems, Fric could identify some weak fricatives as particular phonemes that one expert could only label as weak. This feat was a result of the power of gaining knowledge from multiple sources.

7.4.6 Conclusions

One major advantage of this approach was seen during testing. Many of the template matching solutions have trouble when the segmentation is not precise. Expecting precision from the segmenter is somewhat unrealistic as boundaries between sounds are not often clear. When identifying segments, it was found the boundaries could be changed without effecting the performance of Fric. Since this system looks for specific features, details about the segment can be changed without affecting the identification of the segment. Fric not only functions as a fricative recognizer, but also shows that this approach to phoneme recognition is well worth implementation.

7.5 Phoneme Database

A critical component of the system is a "phoneme" database that stores pairs of intermediate strings and words. Our studies have shown that an average of five phoneme strings is needed for each word in the system's vocabulary to account for difference in pronunciation. This means that even for a moderate vocabulary, a large database is needed to store all the ways of "saying" the words in that vocabulary. The implementation of that database and its retrieval speed, then, are critical in constraining the approaches to be tried for the organization of the word builder, not to mention the performance of the final system.

The work done to date on the design of this database has centered on alternative approaches for its physical layer. Since the only critical function is retrieval, a full blown database management system is neither necessary nor desirable. Therefore, work has focused on random access retrieval strategies like inverted files, indexing and various hashing techniques.

The main problem in matching phoneme strings to words is the "fuzziness" of the phoneme strings. As described in section 7.1.2, a given phoneme string might consist of very certain sounds mixed with uncertain sounds. In its first version, the retrieval module will not deal with this fuzziness, but it will instead be a "dumb but fast" retrieval of a word given a string of exact phonemes.

If retrieval of exact phoneme strings can be done fast enough, the word building module that provides the phoneme strings to look up in the lexicon can be organized around a generate-and-test strategy. This module, then, would hypothesize all possible exact phoneme strings for a given fuzzy phoneme string, look each of them up in the lexicon, and collect the words that matched.

If retrieval speeds become a bottle neck, however, a smarter retriever that deals with uncertain phonemes on a lower level will have to be considered. However, we still anticipate the need for a low level dumb-but-fast retriever to act as the actual interface to the lexicon database. It is not clear yet how close to the actual retrieval module the uncertainty handling mechanisms will have be.

The design being completed at this time for the physical structure of the lexicon is basically an inverted file organization augmented with special hashing techniques to speed up retrieval. We estimate that retrieval speeds on the available hardware will be sufficient to provide a reasonable chance of success for a generate-and-test word builder module.

BIBLIOGRAPHY

Catford, J.C. "The articulatory possibilities of man", In B. Malmberg (Ed.) Manual of Phonetics, Amsterdam, Holland: North Holland Publishing Co., 1957, 309-333.

Cole, R.; Rudnicky, A.; Zue, V. and Reddy, D.R. "Speech as Patterns on Paper", In R.A. Cole (Ed.) Perception and Production of Fluent Speech, Hillsdale, NJ: Lawrence Erlbaum Associates, 1980, 3-50.

Denes, P.B. "Automatic speech recognition: Old and new ideas", In D.R. Reddy (Ed.) Speech Recognition, NY: Academic Press, 1975, 73-82.

Fant, G. "Analysis and Features of Speech Processes", In B. Malmberg (Ed.) Manual of Phonetics, Amsterdam, Holland: North Holland Publishing Co., 1957, 173-273.

Fant, G. Speech Sounds and Features, Cambridge, MA: MIT Press, 1973.

Francis, W.N. The Structure of American English, NY: Ronald Press Co., 1958.

Gerstman, L. "Classification of self-normalized vowels", IEEE Transactions on Audio and Electroacoustics, AU-16, 78-80.

Glass, R., "Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment," M.S. Thesis, MIT, 1984.

Glass, J. and Zue, V. "Recognition of nasal consonants in American English", Proceedings of the DARPA Speech Recognition Workshop, Palo Alto, California, February, 1986.

Hecker, M.H.L. "Speaker recognition - An interpretive survey of the literature" American Speech and Hearing Association, Washington D.C., Jan., 1971.

Hermansky, H., Hanson, B., and Wakita, H. "Perceptually based analysis of speech", Proc. ICASSP-85, 1985, 509-512.

Klatt, D. "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", Journal Acoustical Society of America, 59, 1976, 1208-1221.

Klatt, D. "Speech perception: A model of acoustic-phonetic analysis and lexical access", In R.A. Cole (Ed.) Perception and Production of Fluent Speech, Hillsdale, NJ: Erlbaum, 1980.

Klatt, D. and Stevens, K.N. "On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment," IEEE Trans. Audio and Electroacoustic, Vol.

AU-21, June 1973, 210-217.

Kopec, G., "Formant Tracking Using Hidden Markov Models and Vector Quantization," IEEE Trans. Acoust. Speech, Signal Proc., Vol. ASSP-34, August 1986.

Markel, J. and Gray, A.H. Linear Prediction of Speech, New York: Springer-Verlag, 1976.

McCandless, S., "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust., Speech, Signal Proc., vol. ASSP-24, April 1974. p-4 Miller, J.D. "A phonetically relevant auditory-perceptual space", Journal of the Acoustical Society of America, 72, (Suppl. 1), S64 (A) 1982.

Miller, G.A. and Nicley, P.E. "An analysis of perceptual confusions among some English consonants", Journal of the Acoustical Society of America, 27, 1955, 338-352.

Minifie, F.D. "Speech acoustics", In F.D. Minifie, T.H. Hixon, and F. Williams (Eds.), Normal Aspects of Speech, Hearing, and Language, Englewood Cliffs, NJ: Prentice Hall, 1973, 235-284.

Niederjohn, R. and M. Lahat, "A zero-crossing consistency method for formant tracking of voiced speech in high noise levels," IEEE Trans. Acoust. Speech, Signal Proc., vol. ASSP-33, April 1985.

Peterson, G. and Barney, H. "Control methods used in a study of the vowels", Journal of the Acoustical Society of America, 24, 1952, 175-184.

Potter, R., Kopp, G.A., and Green, H.G. Visible Speech, NY: Dover Publications, 1966.

Radian Corporation RuleMaster Reference Manual, Radian Corp., Austin, Texas, Nov. 1985.

Reise, C.E. and Zubrick, S.M. "RuleMaster - Using Induction to Combine Declarative and Procedural Knowledge Representations," Radian Technical Report R1-R-00299, 1985.

Seneff, S., "Pitch and spectral analysis of speech based on an auditory synchrony model," Ph.D. dissertation, MIT, 1985.

Shepard, R.N. Multidimensional scaling, tree-fitting, and clustering", Science, 210, 1980, 390-398.

Zue, V. "Notes on Spectrogram Reading", Preliminary Draft, M.I.T., 1985.

Zue, V. "The Use of Speech Knowledge in Automatic Speech

Recognition". Proceedings of the IEEE. Vol. 73, no. 11, Nov. 1985, 1602-1615.

Zue, V. and Cole, R.A. "Experiments on Spectrogram Reading", ICASSP 79 Record.